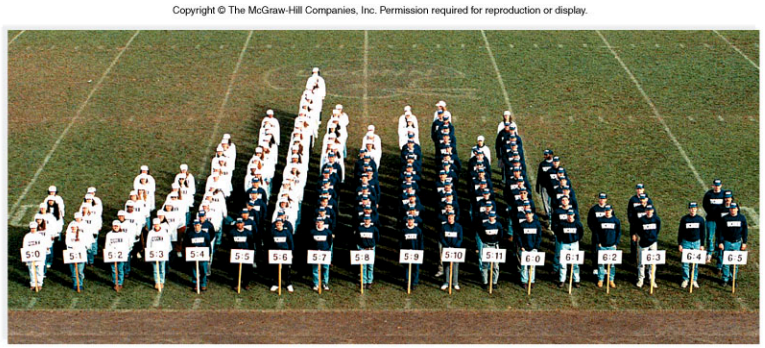
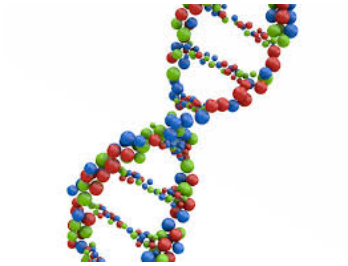
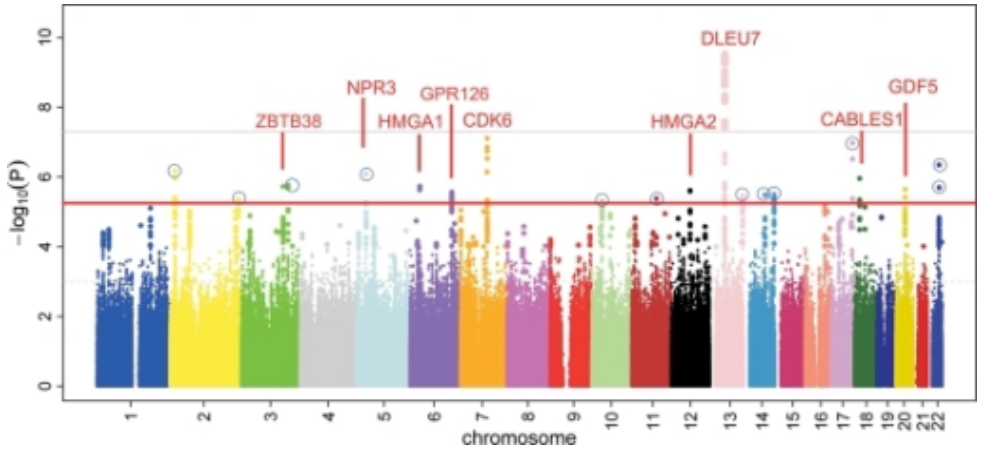
Tips in Data Visualization for Genetic Mapping







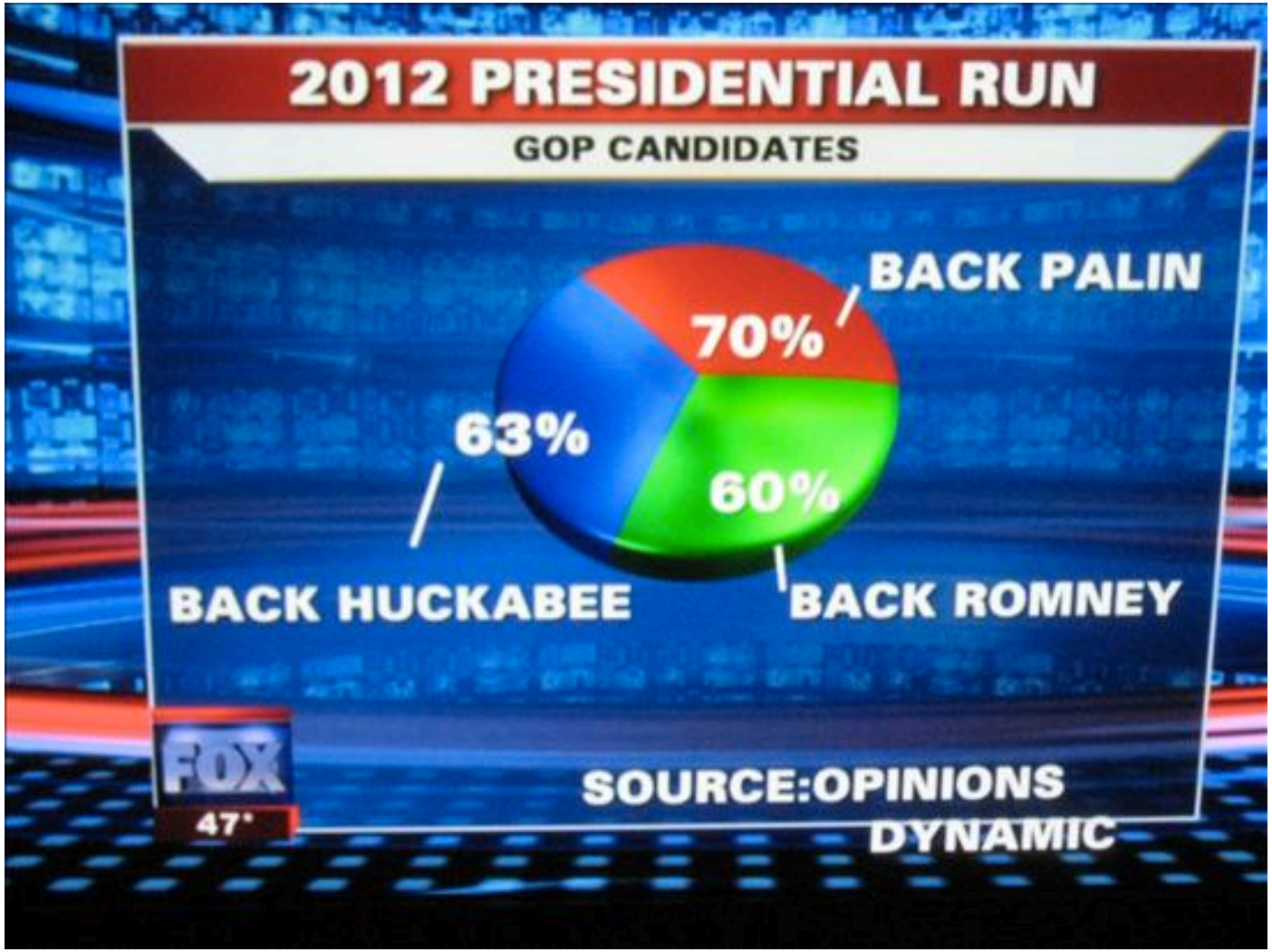
Jenn Coughlan + Ryan Campbell

Learning goals:

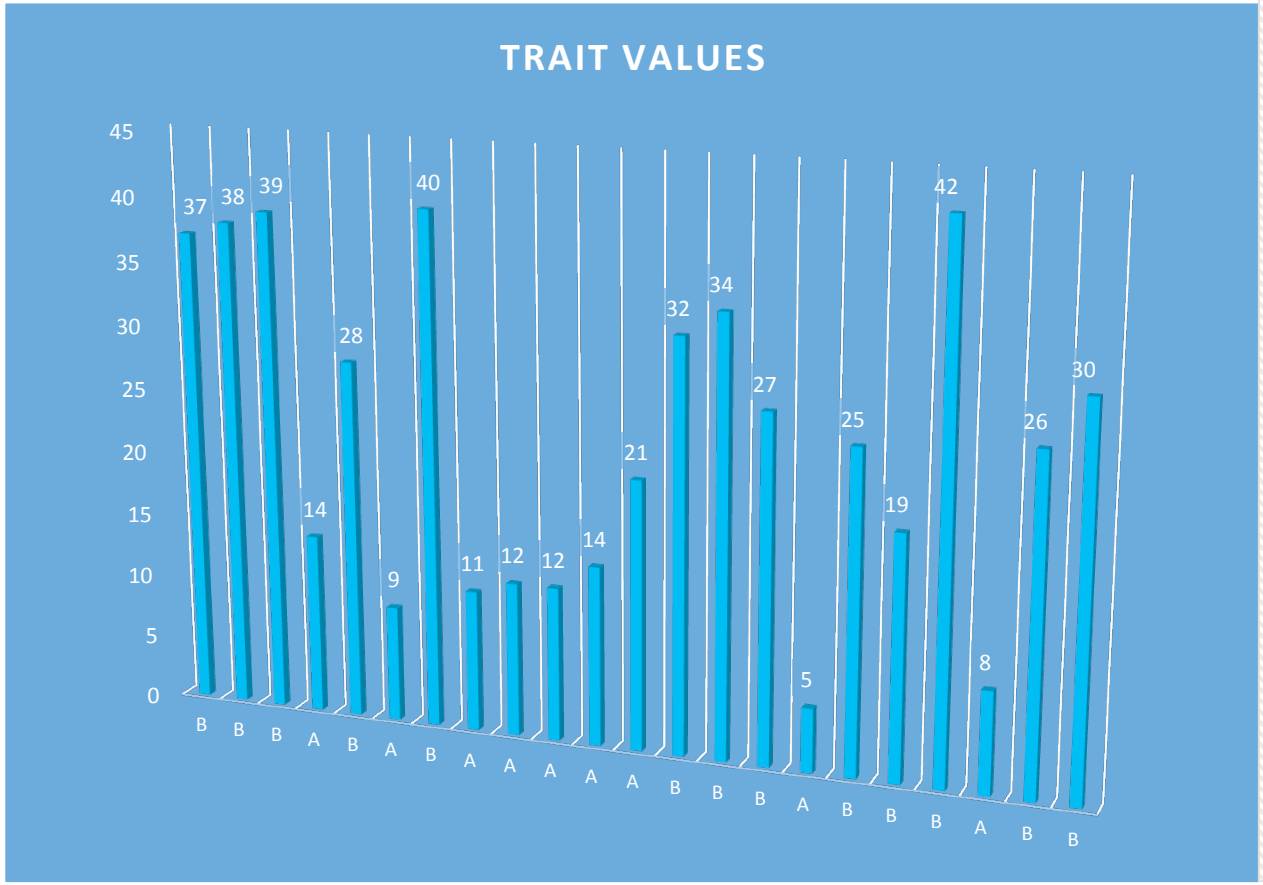
- Be able to list types of + explain the concept of genetic mapping
- Compare/contrast multiple tools for plotting graphs
- Produce a figure using base plot and ggplot functions
- Manipulate big data to make a figure with complicated data

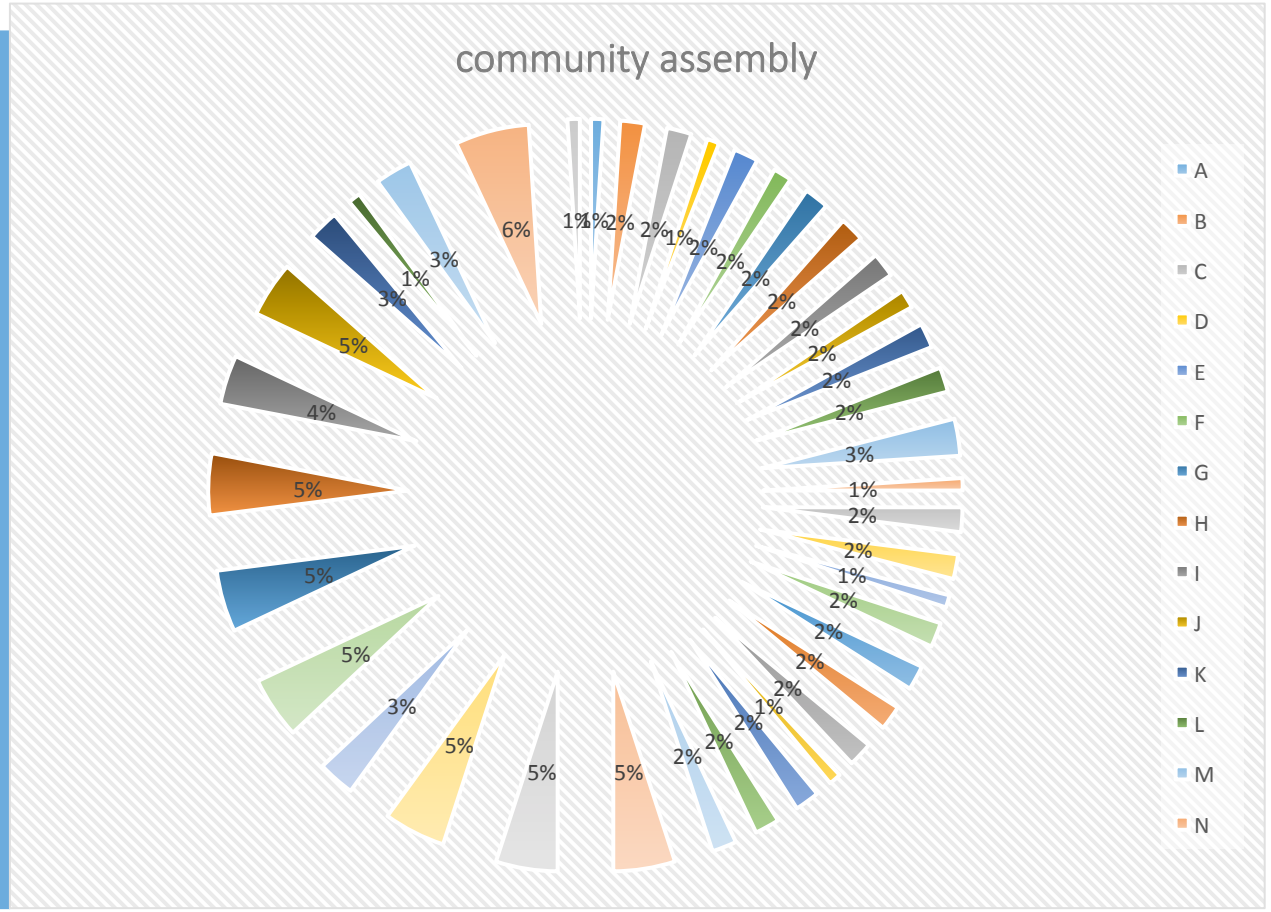
What is data visualization? Why is it useful?

Sometimes, graphs are confusing:



How can graphs be bad? Inappropriate type of graph, too complicated, poor aesthetic choices

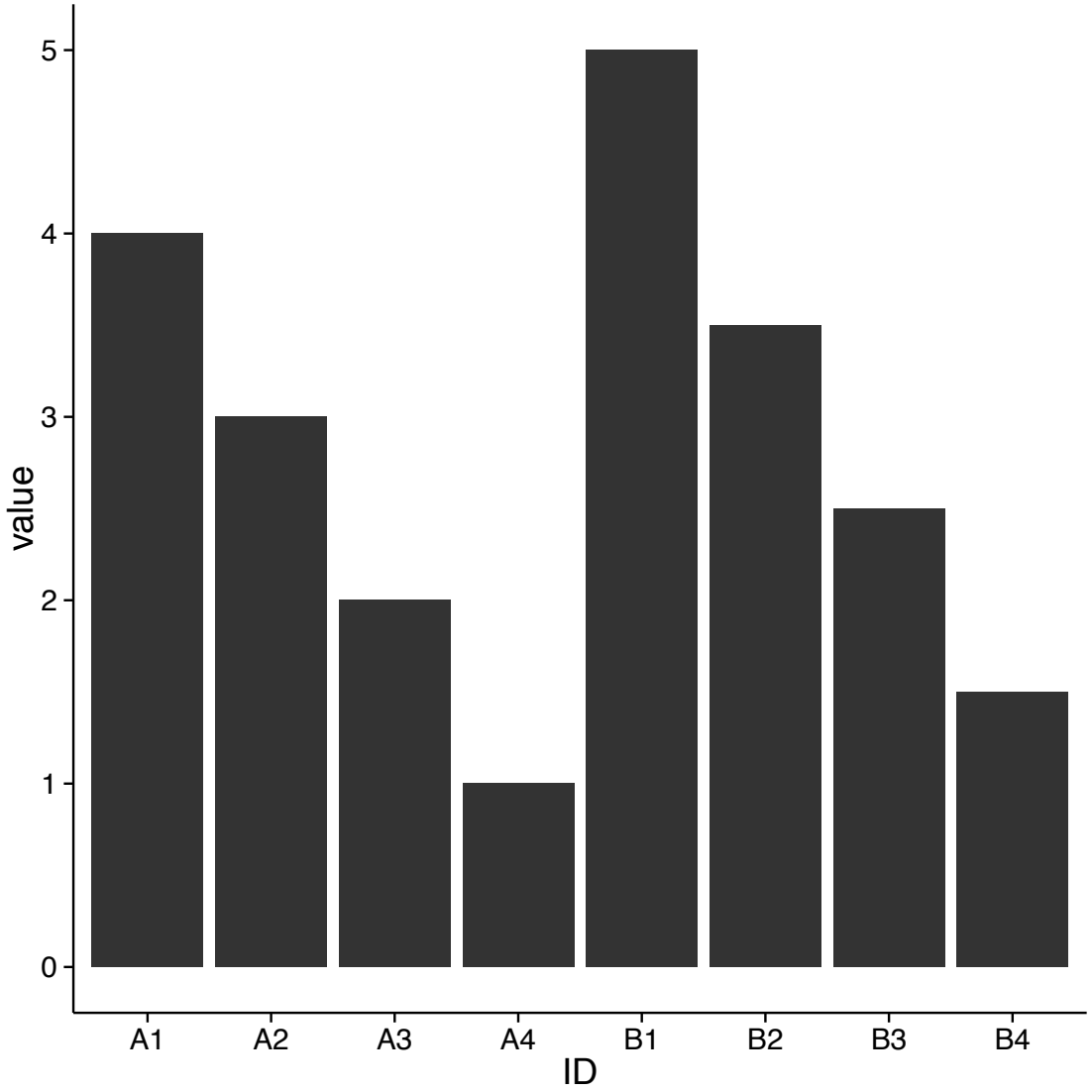




Tips on making nice figures

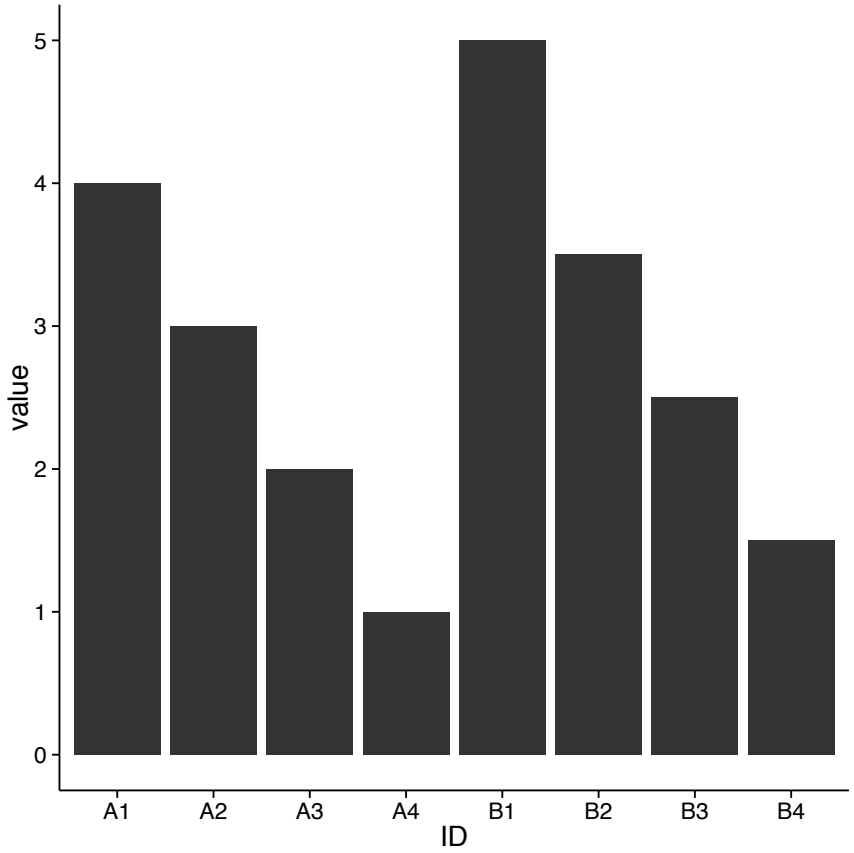
- What is the story you want to tell with the data? Plot data to emphasize that.
- Be mindful of the type of graph you use
- Think about color choice (and other design options)
- Simple is usually better

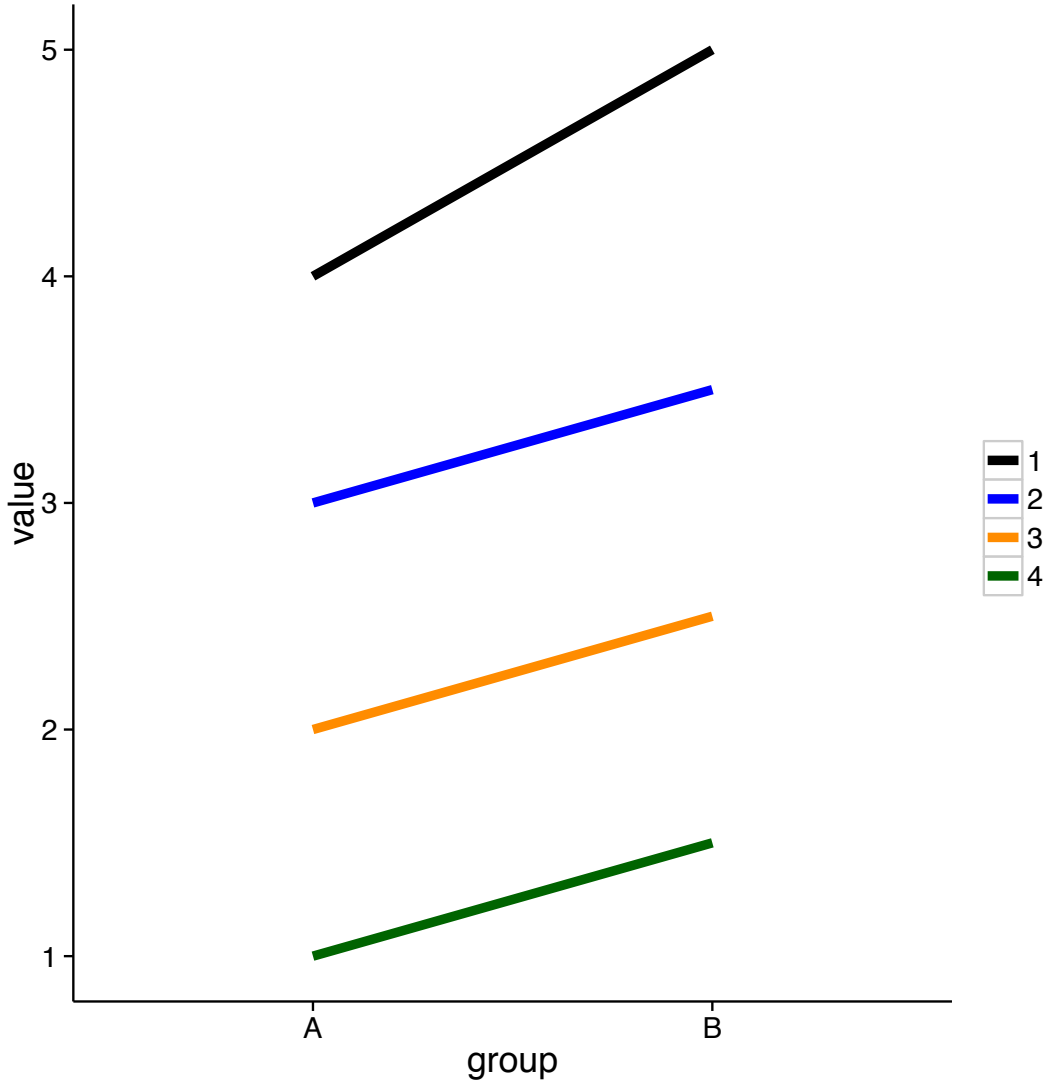
What is the story you want to tell with the data?



Borrowed from Collin Maxwell: guest lecture in Bio 760

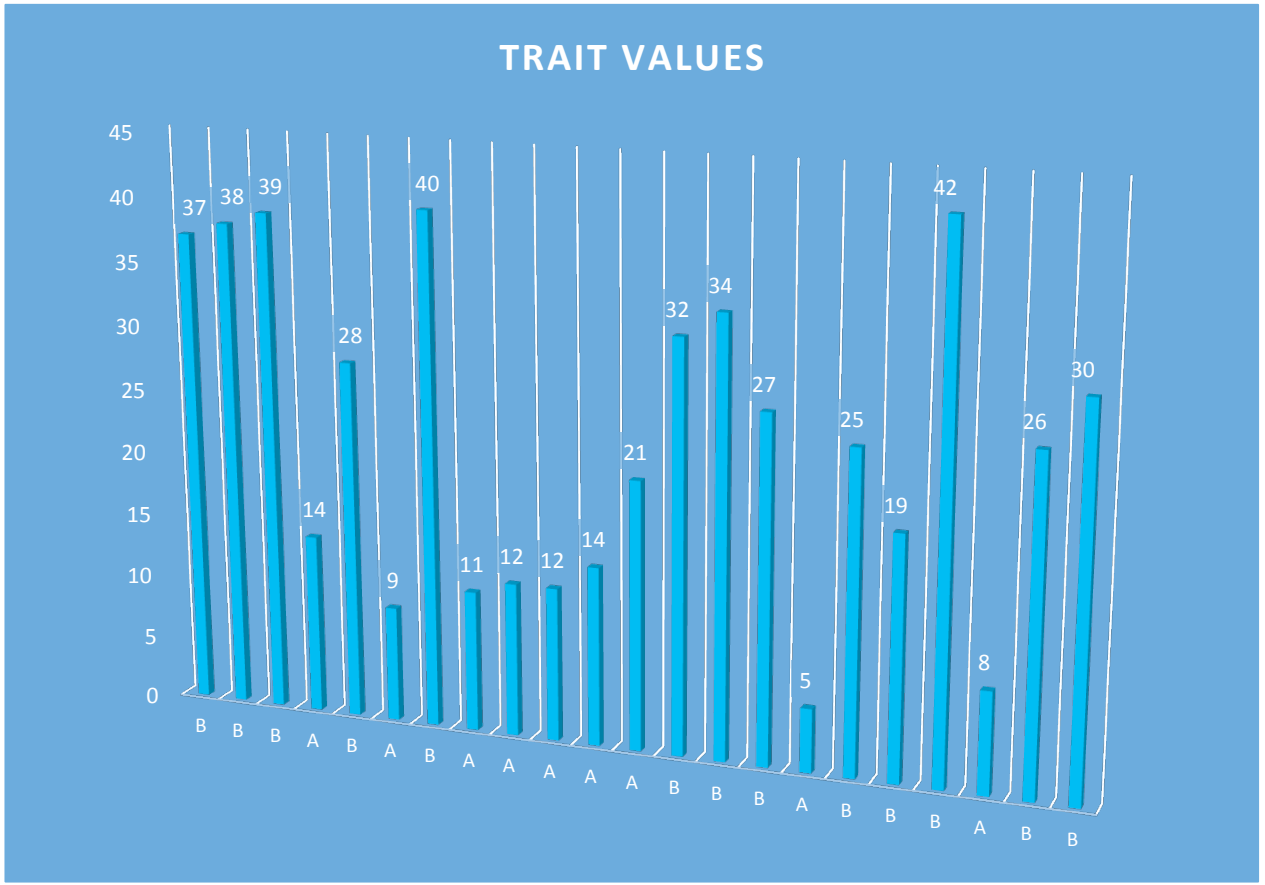
What is the story you want to tell with the data?



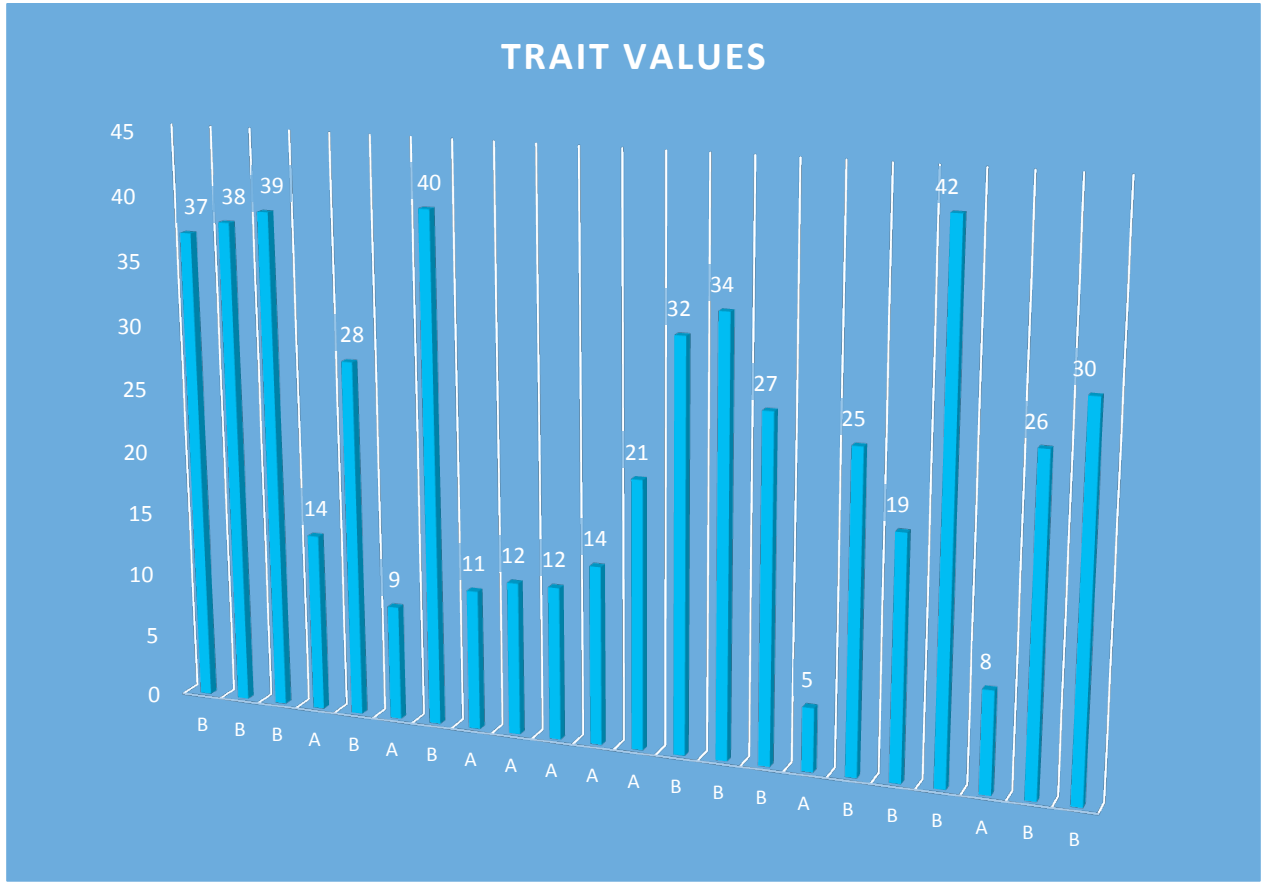


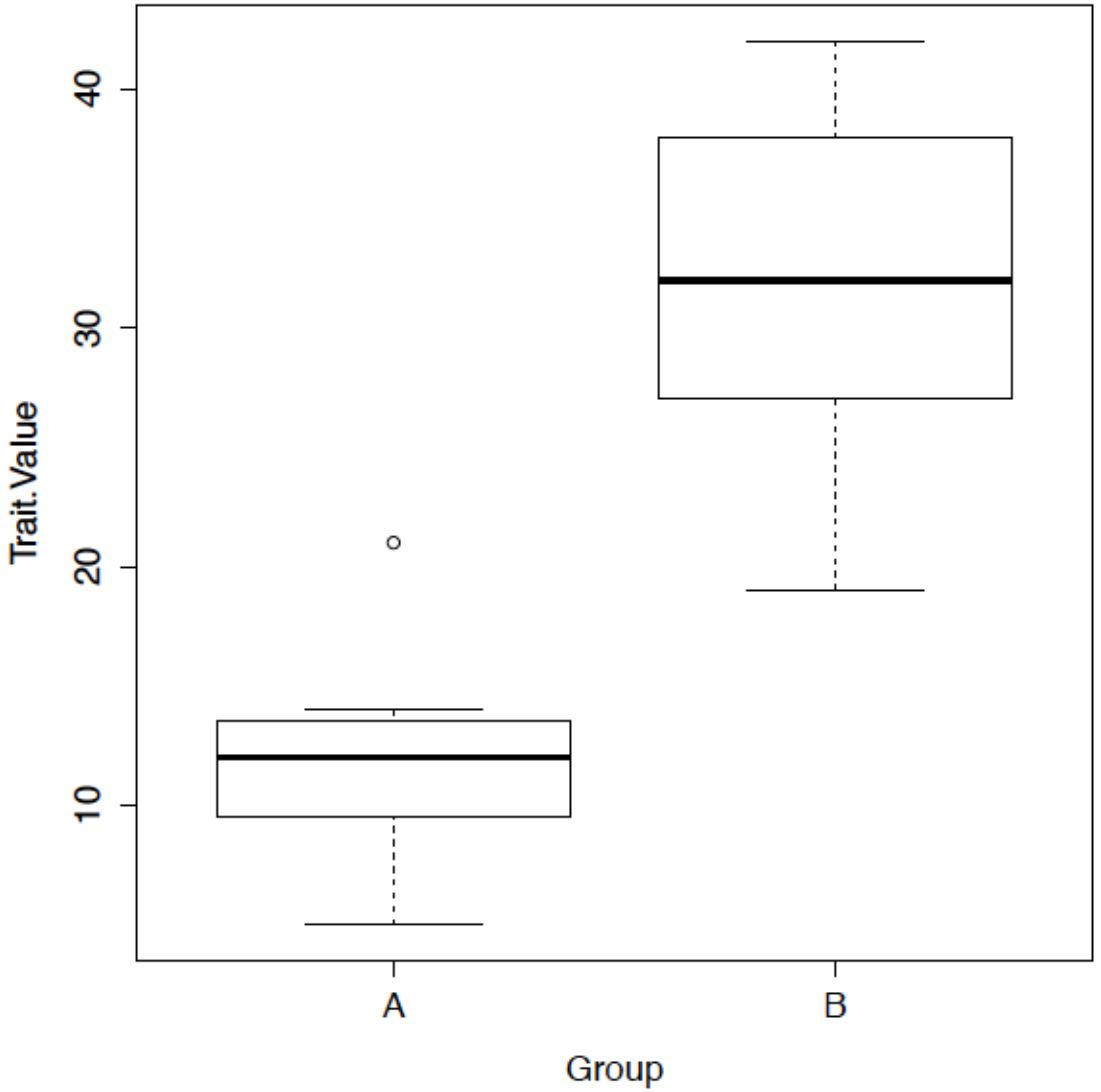
Borrowed from Collin Maxwell: guest lecture in Bio 760

Be mindful of the type of graph you use

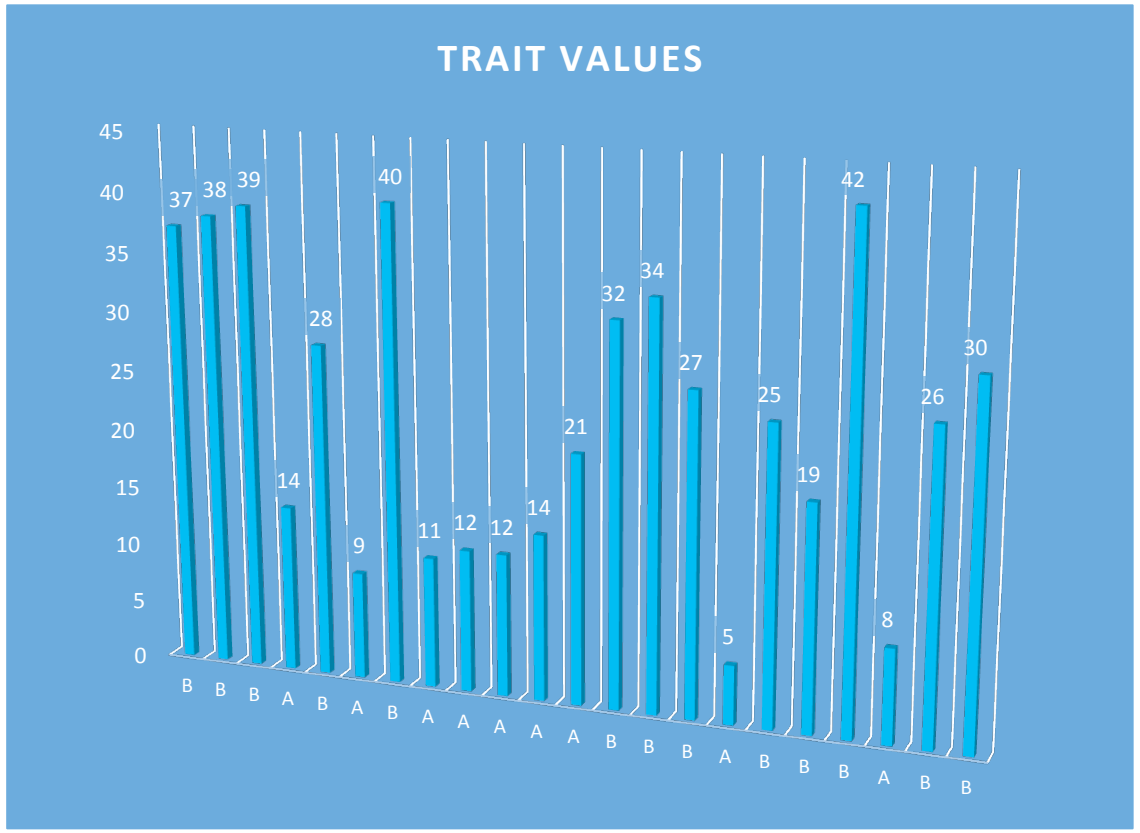


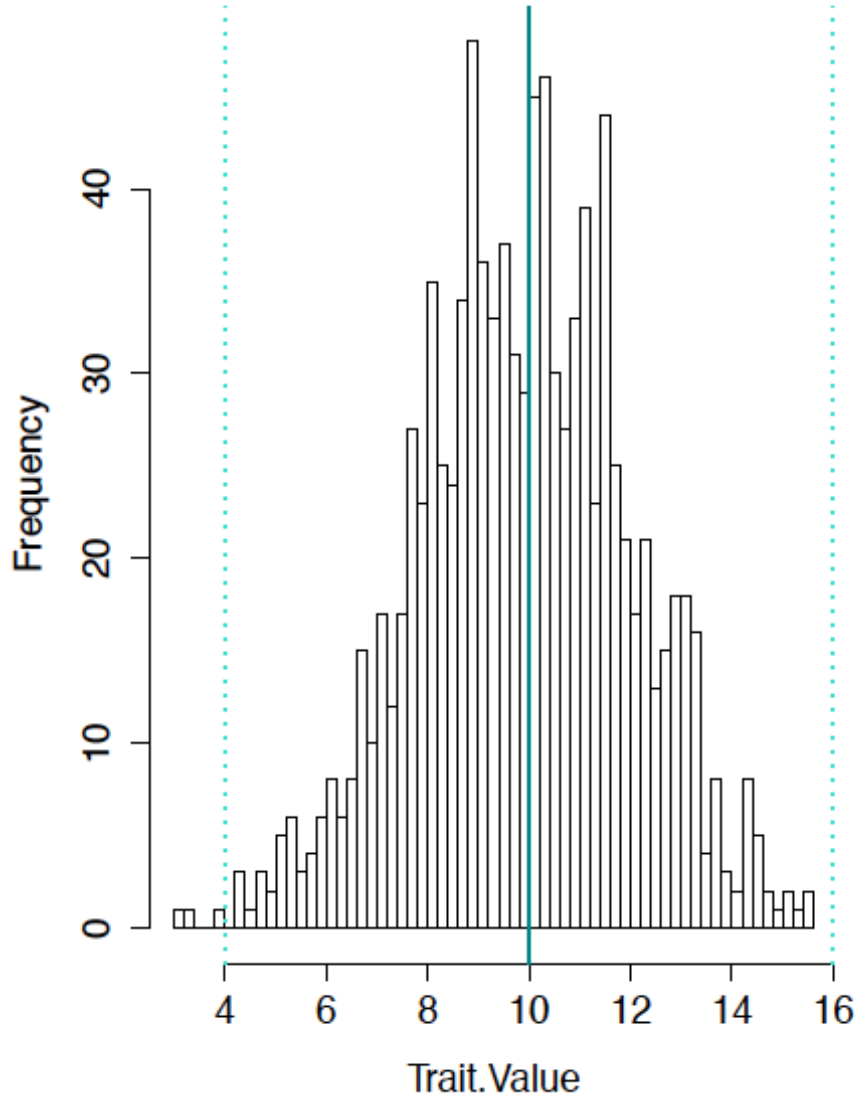
Be mindful of the type of graph you use



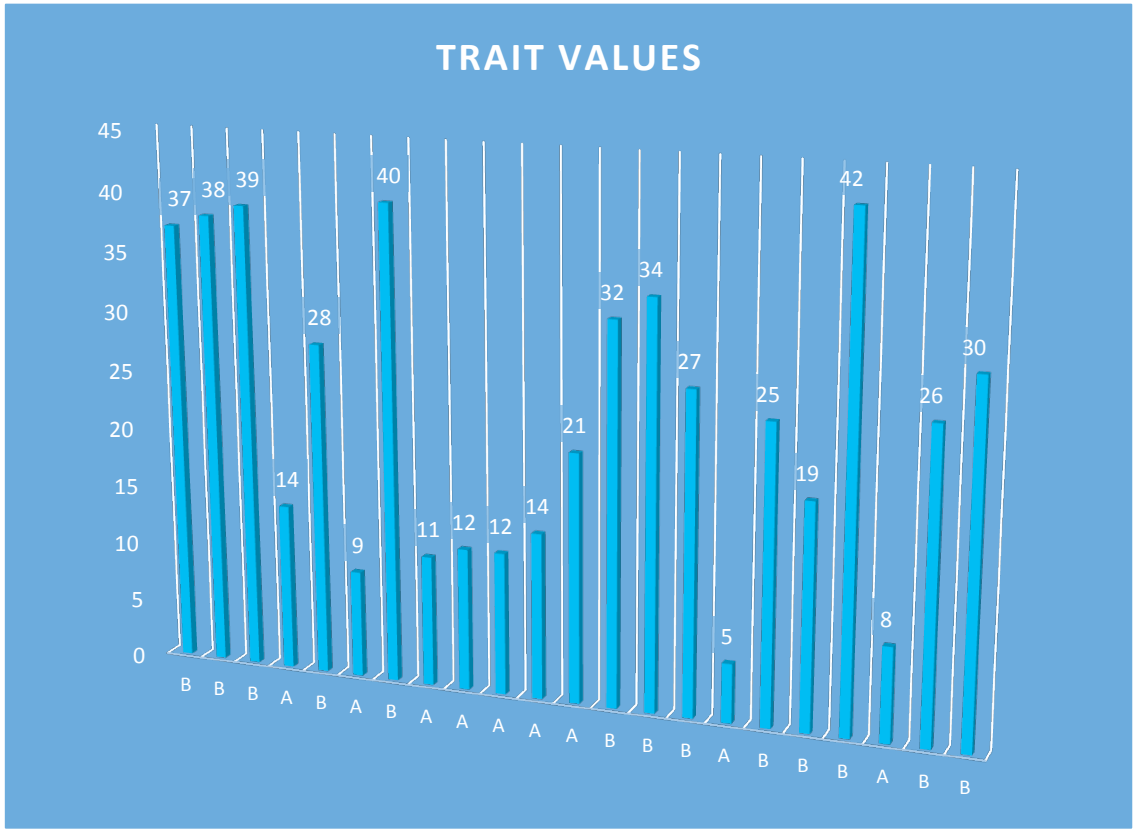


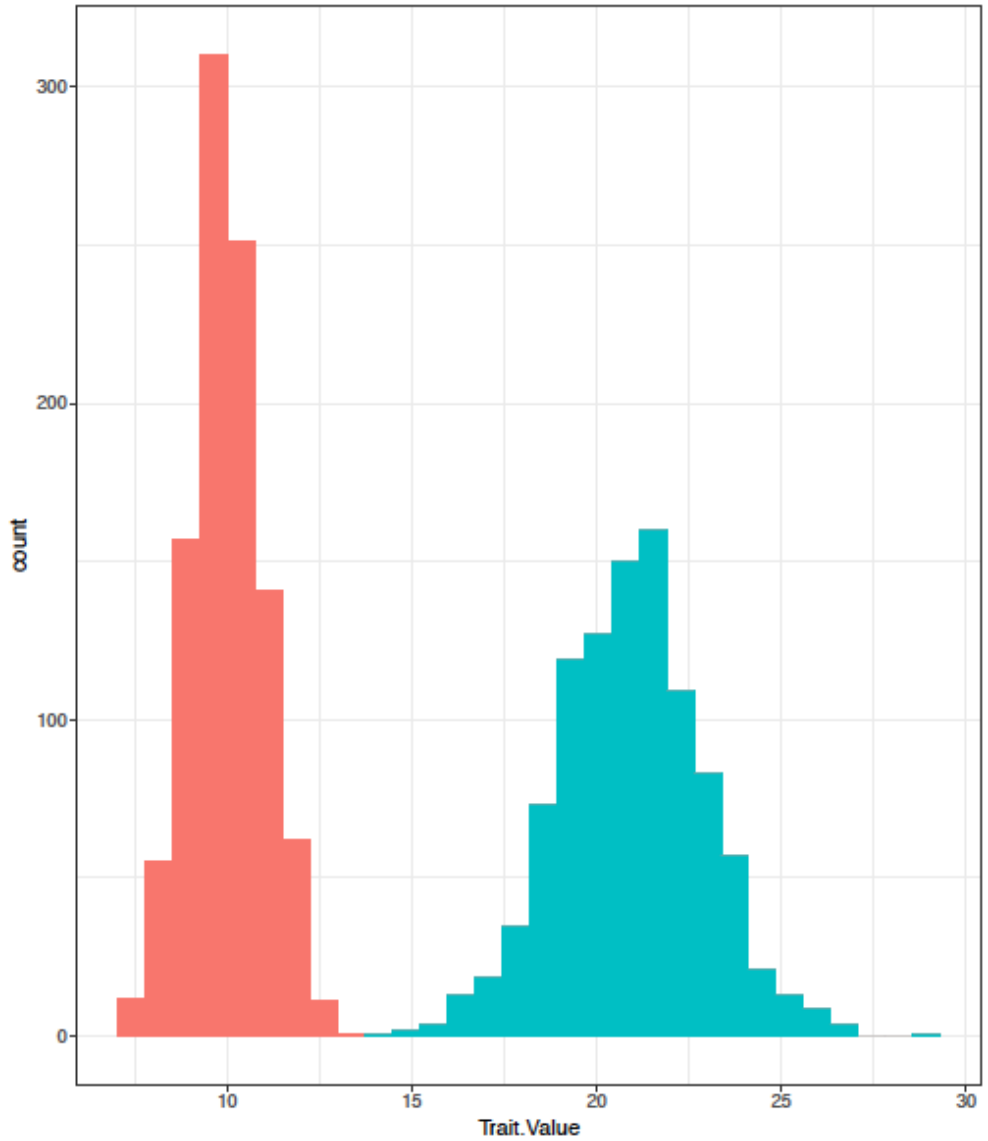
Be mindful of the type of graph you use





Be mindful of the type of graph you use

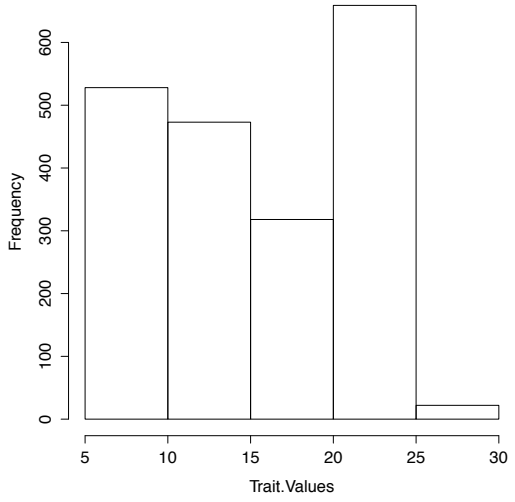


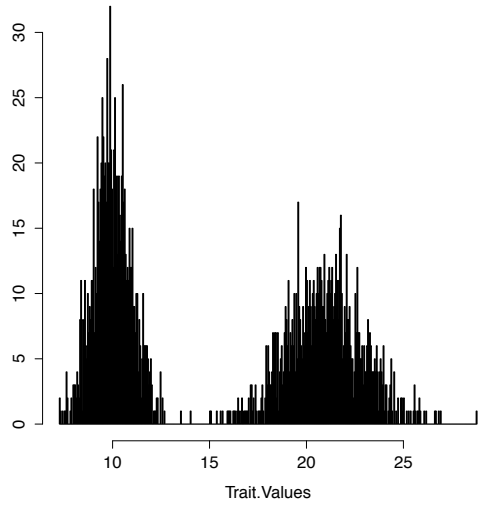


Think about design

- Can create colour palettes, or use pre-defined ones (i.e. RColorBrewer, Wes Anderson)
  - Think to think about: contrast, color wheels(!), how colors will print in black and white, and color-blind friendly colors
- Making font readable
- Background design(R base plot uses white, ggplot uses grey with lines)
- Other design choices (i.e. bin size for histograms, notches on boxplots)

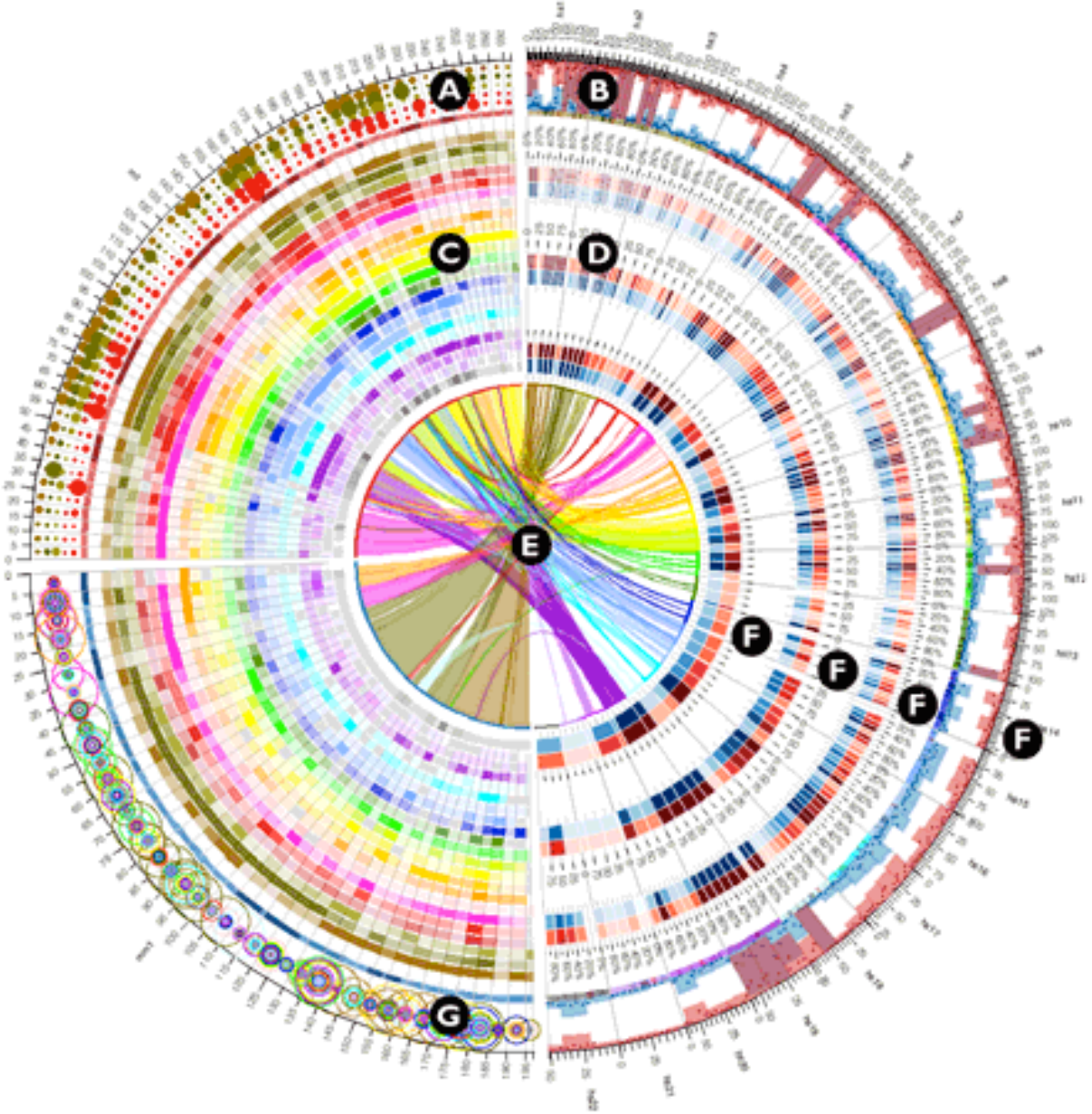
R 'Wes Anderson' colour palettes





Colors in R: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

Simple is often better



Tools for making nice figures

- **Excel:** easy, sometimes clumsy, some ability to change aesthetics, only simple statistics, user friendly

- **JMP:** easy, sometimes clumsy, limited ability to change aesthetics, but can do [slightly more complicated] statistics

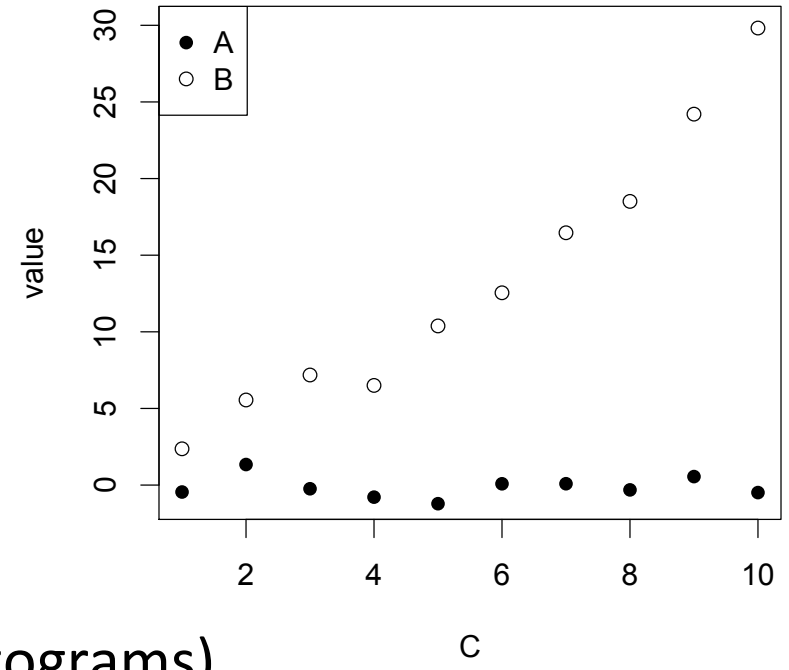
- **Adobe Illustrator:** Exceptional ability to change aesthetics, not user friendly, no statistics ability

- **R:** exceptional ability to change aesthetics, requires coding knowledge (but A LOT of tutorials, help online), exceptional, personalizable statistics capability

Specifics of R: Base Plot()

- **plot**(), **boxplot**(), **barplot**(), **hist**() will produce different plot types

- Codes to change different aesthetics:
  - pch→ symbol type
  - col, bg → color and 'fill'
  - lty → line type (full, dashed etc)
  - lwd → line thickness
  - cex → size (for font and symbols)
  - Label functions: (xlab, ylab, main)
  - Specialized aesthics: notch (boxplots), breaks (histograms)



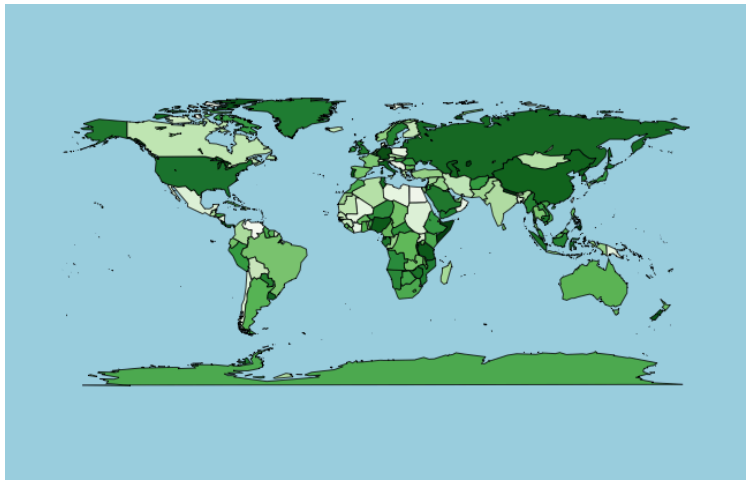
Nice tutorial of base plot:
http://rstudio-pubs-static.s3.amazonaws.com/7953_4e3efd5b9415444ca065b1167862c349.html

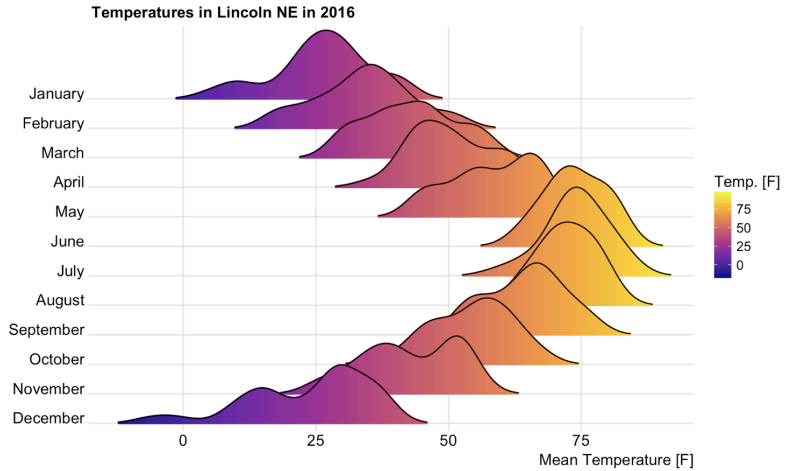
Nice ref for plotting parameters:
http://www.statmethods.net/advgraphs/parameters.html

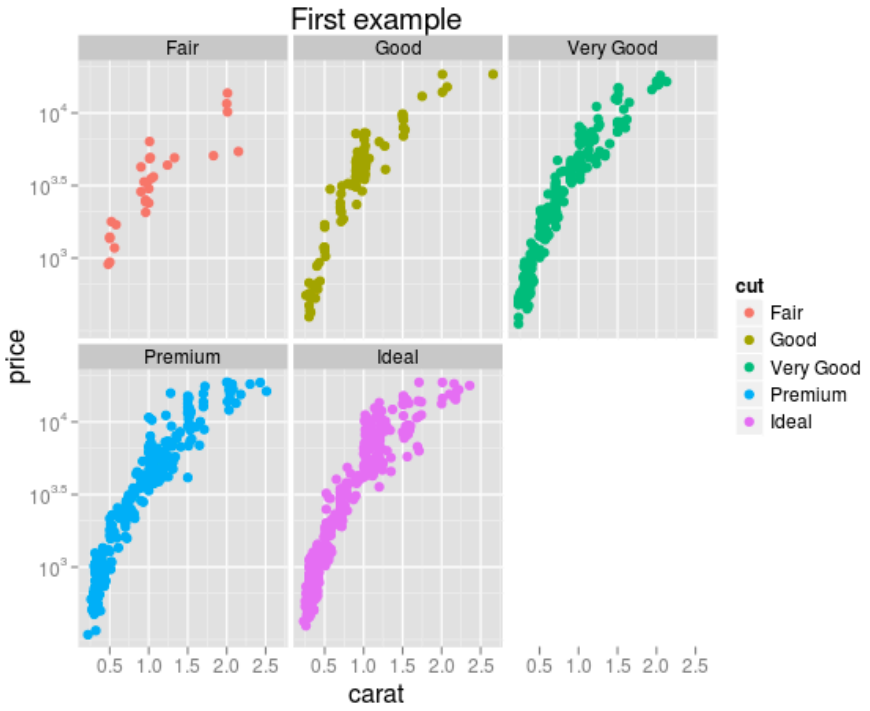
Borrowed from Collin Maxwell: guest lecture in Bio 760

Other (more complicated) figures you can make with R:

- Facets
- Maps
- Animated figures!







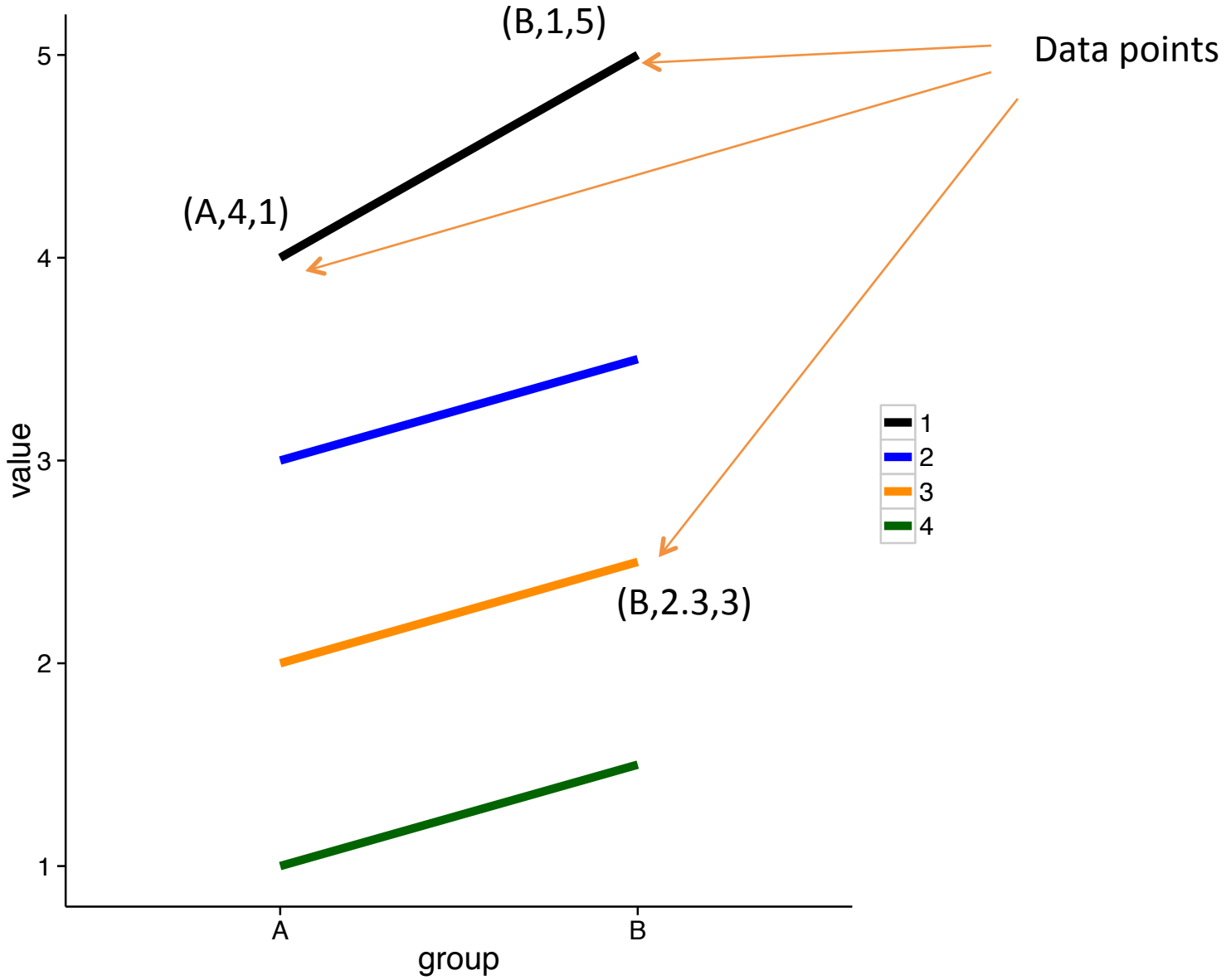
Basic formula of ggplot

ggplot(*dataset*, aes(aesthetic options: x, y, color, size, etc) + **geom_type**(aesthetic options for geom) + **other_aesthetic_decisions**(aesthetics)

Useful overview of geoms and aesthetics:
http://www.cookbook-r.com/Graphs/

Anatomy of a plot

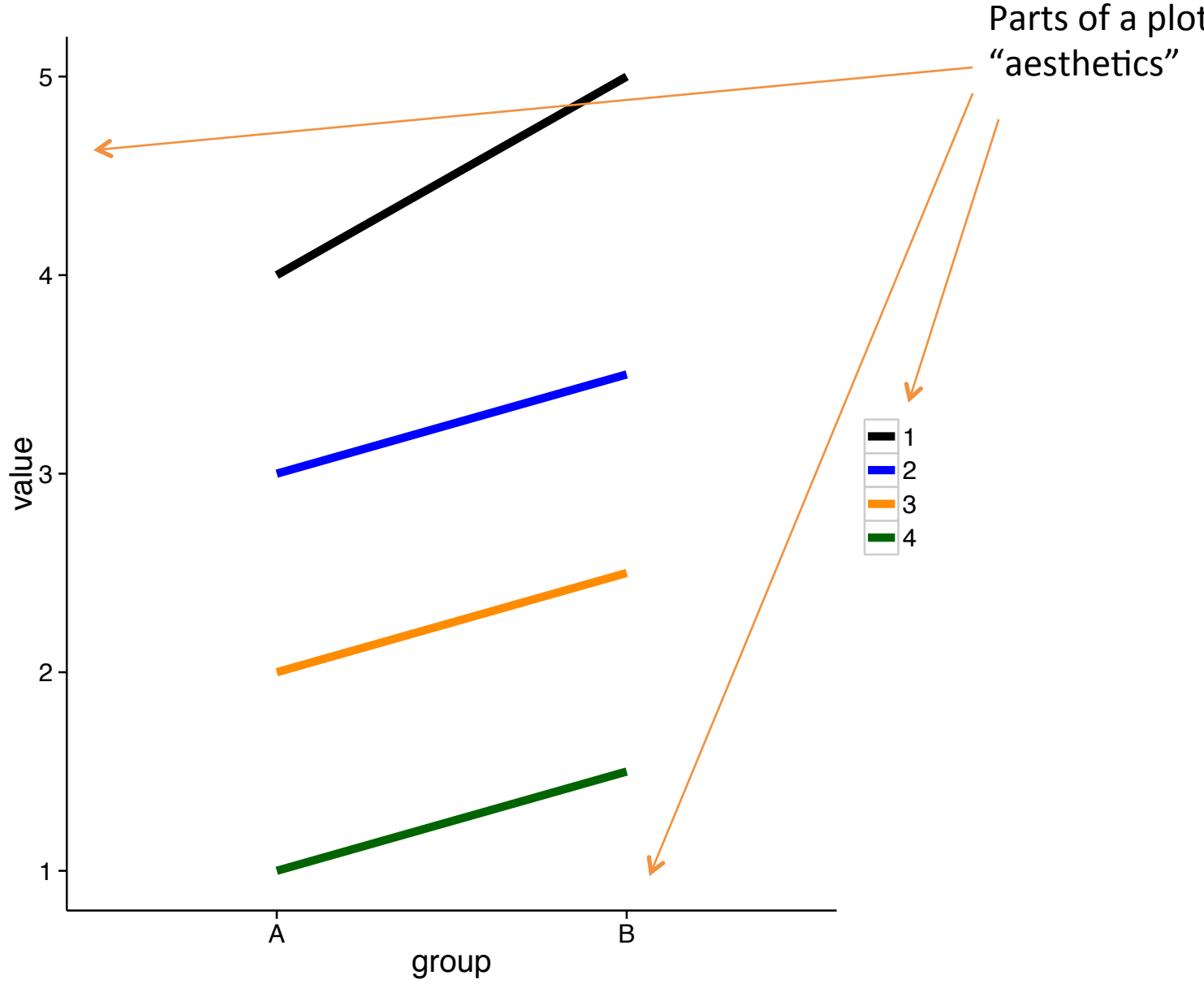
ggplot(*dataset*, aes(aesthetic options: x, y, color, size, etc) +
geom_type(aesthetic options for geom) +
other_aesthetic_decisions(aesthetics)



Borrowed from Collin Maxwell: guest lecture in Bio 760

Anatomy of a plot

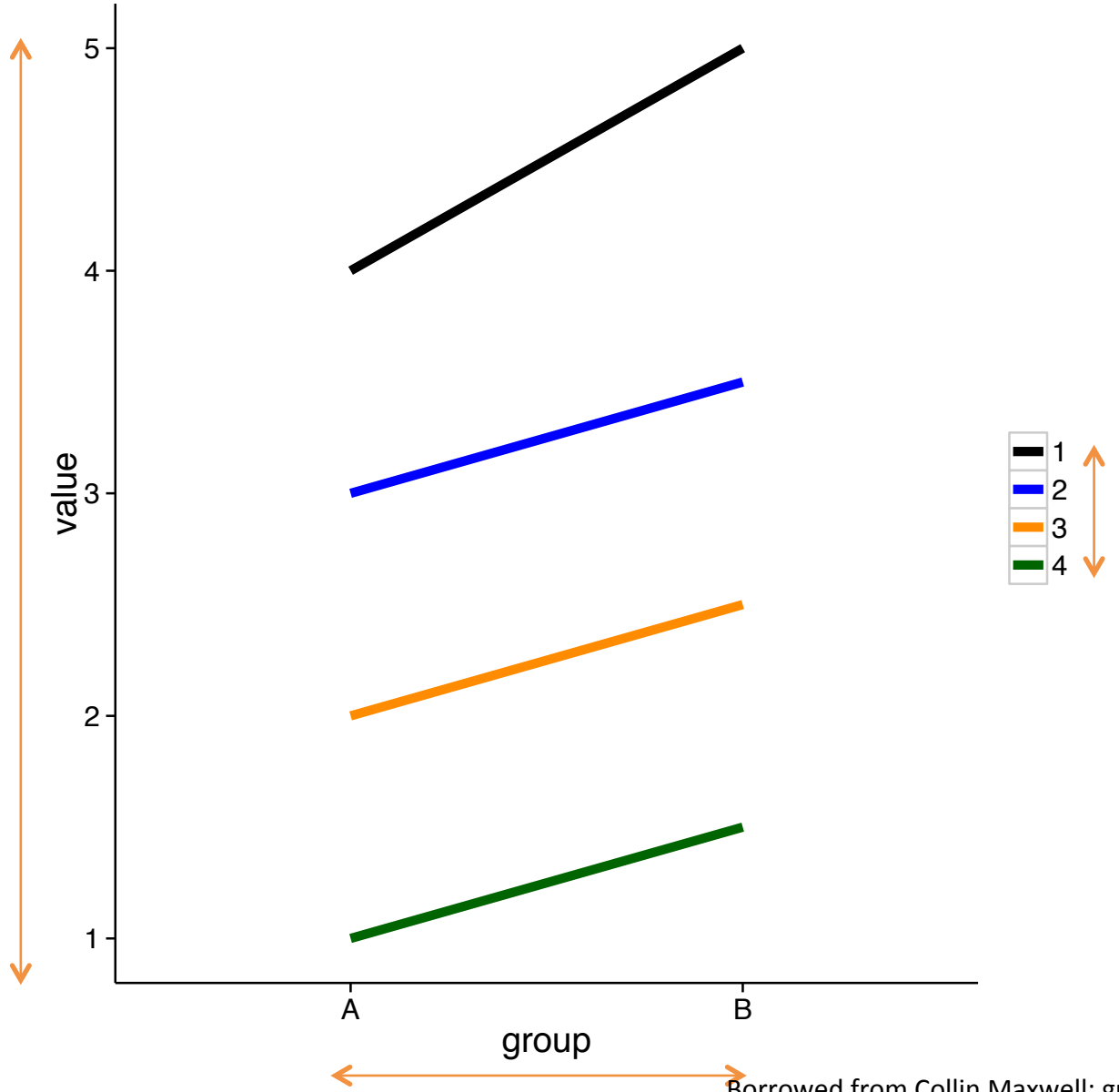
ggplot(*dataset*, aes(aesthetic options: x, y, color, size, etc) +
geom_type(aesthetic options for geom) +
other_aesthetic_decisions(aesthetics)



Borrowed from Collin Maxwell: guest lecture in Bio 760

Anatomy of a plot

ggplot(*dataset*, aes(aesthetic options: x, y, color, size, etc) +
geom_type(aesthetic options for geom) +
other_aesthetic_decisions(aesthetics)

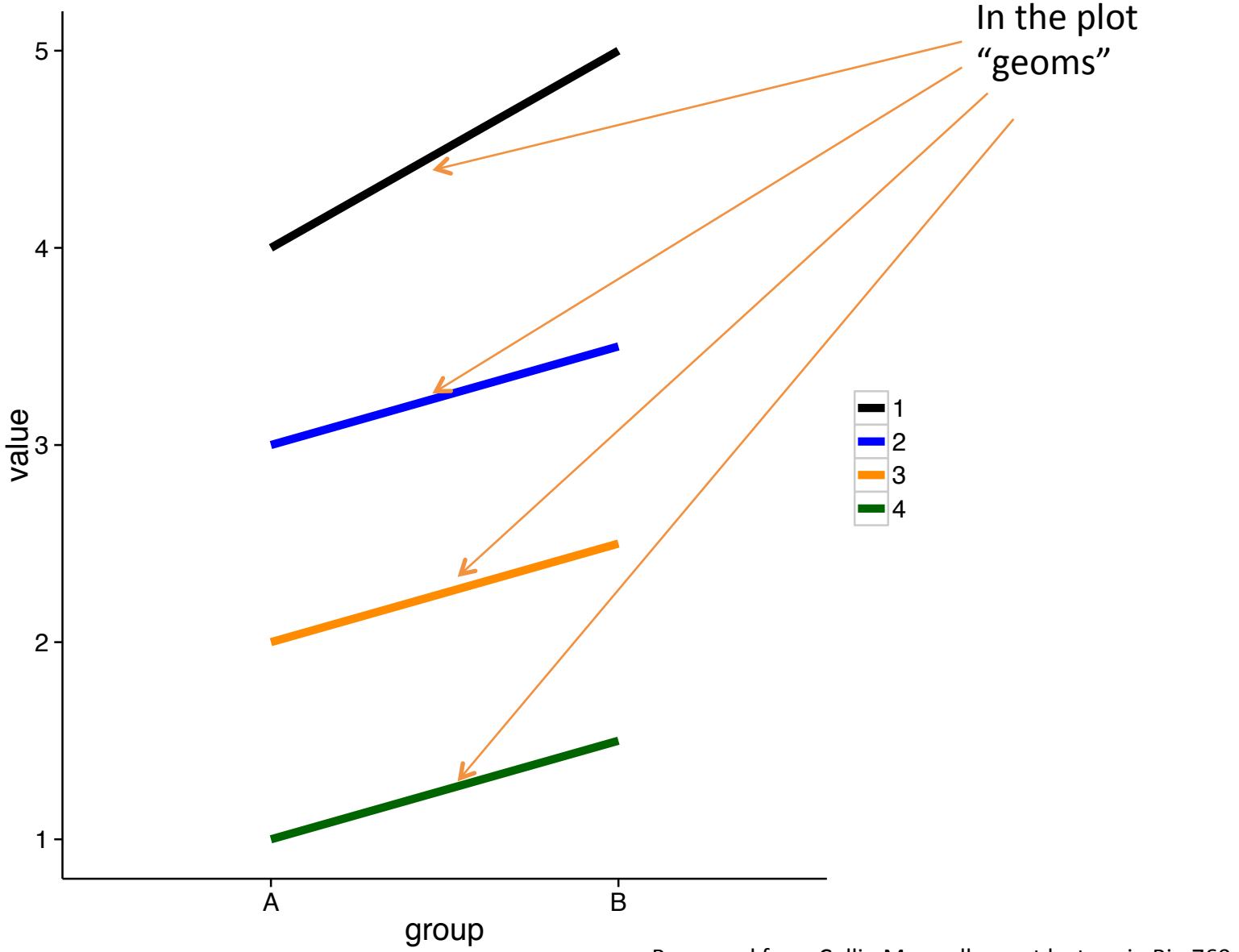


Range of aesthetics "scales"

Borrowed from Collin Maxwell: guest lecture in Bio 760

Anatomy of a plot

ggplot(*dataset*, aes(aesthetic options: x, y, color, size, etc) +
geom_type(aesthetic options for geom) +
other_aesthetic_decisions(aesthetics)



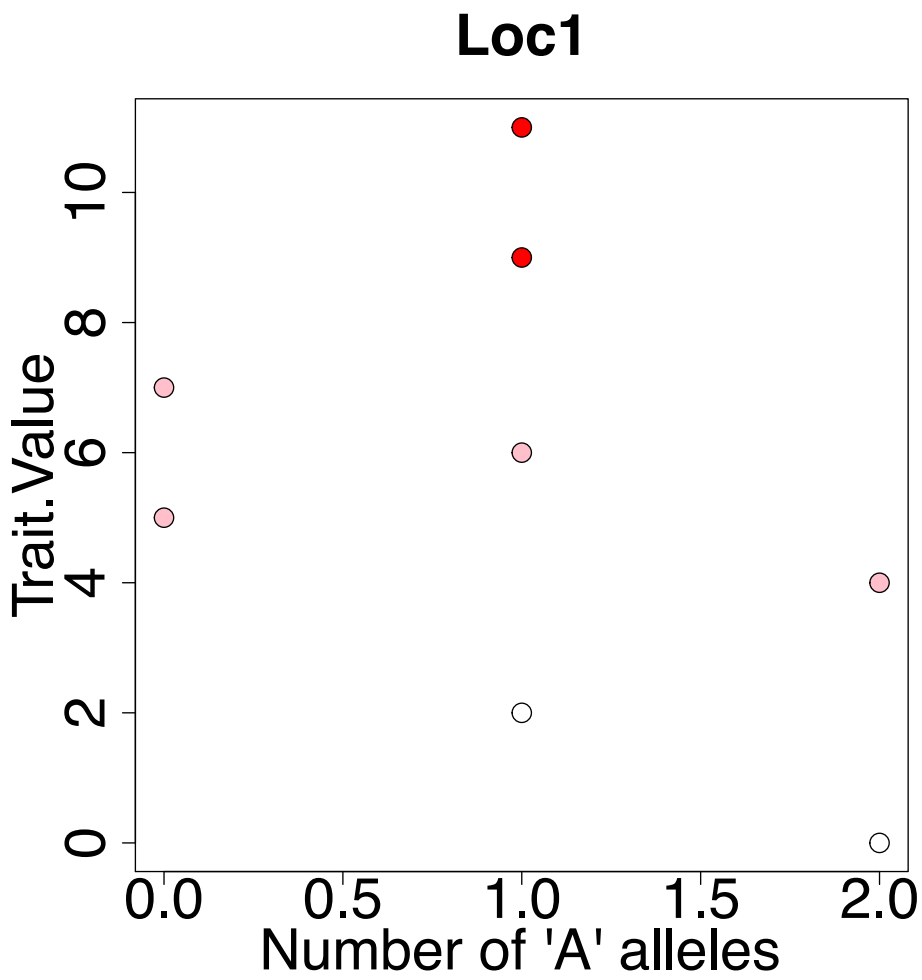
Borrowed from Collin Maxwell: guest lecture in Bio 760

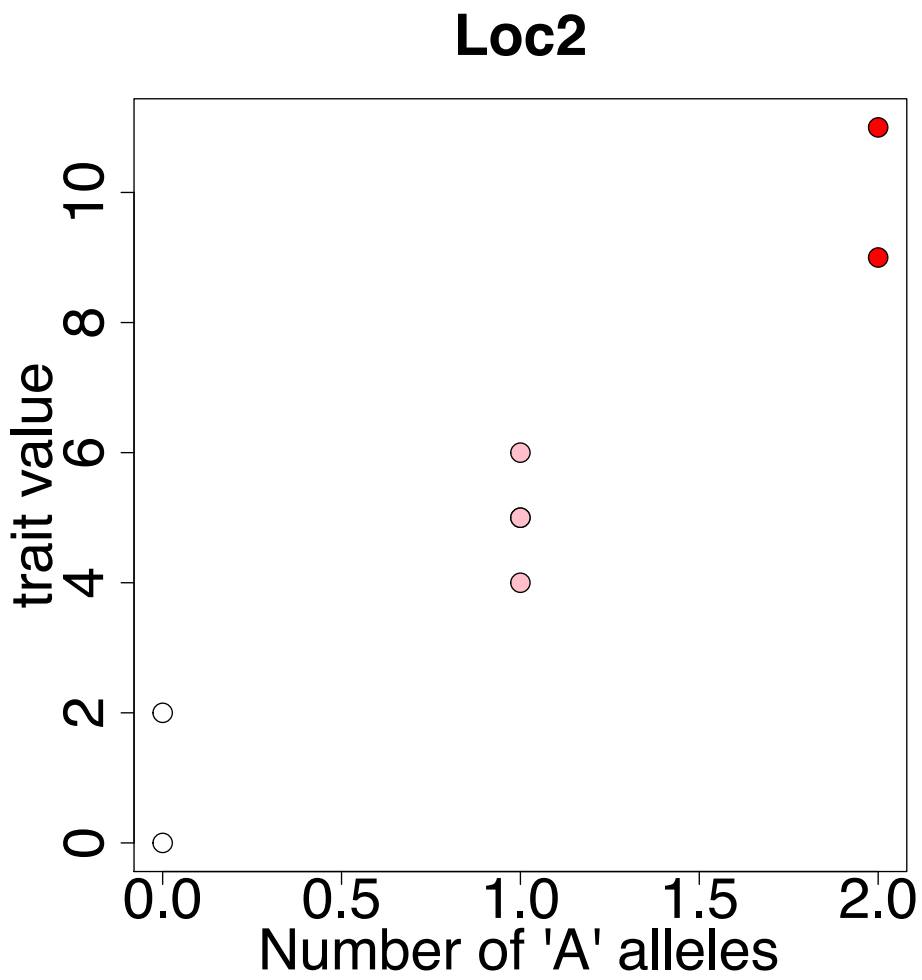
Intro to ggplot2- activity!

What is genetic mapping?

Associating genotype and phenotypes:

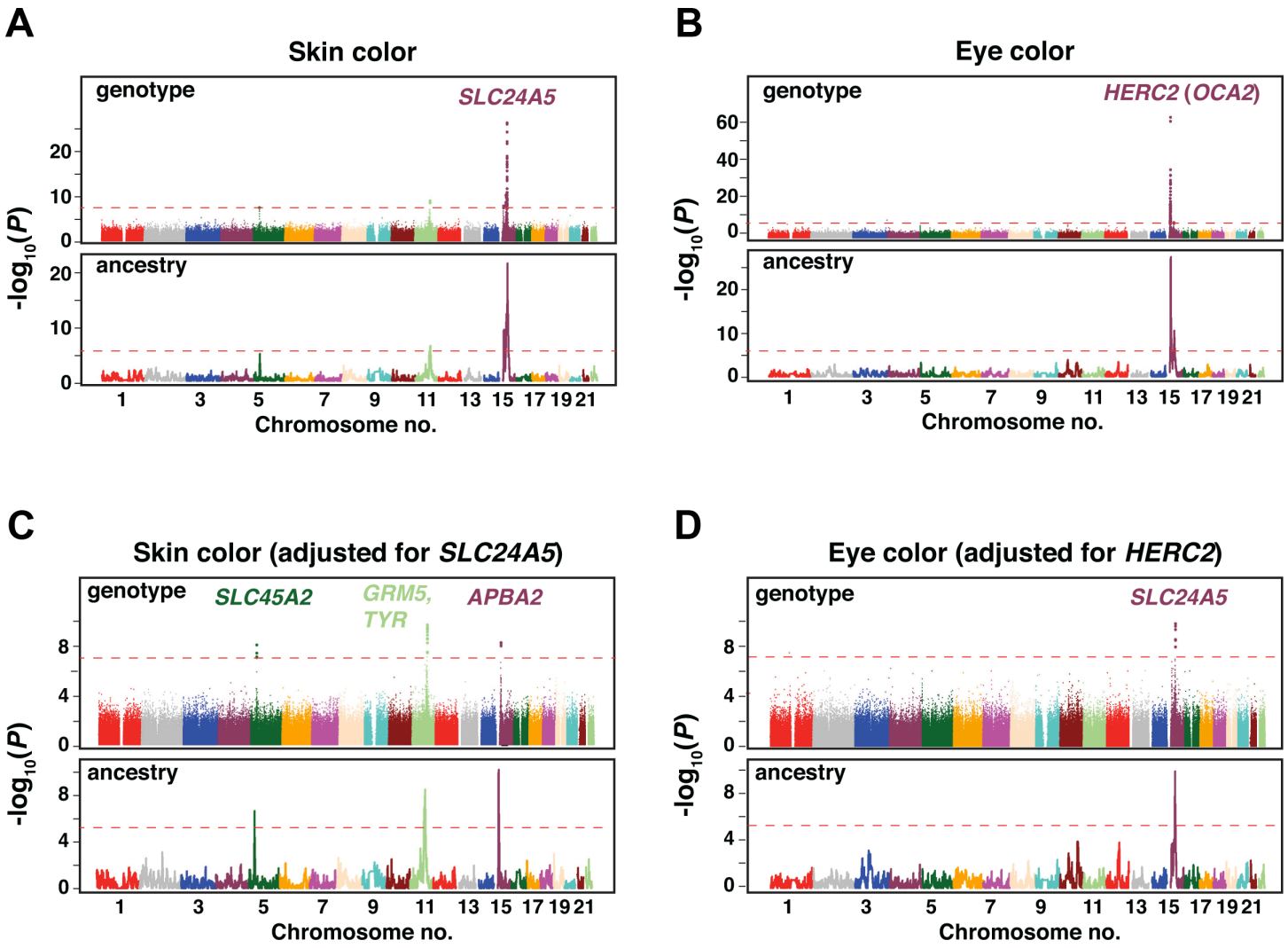
Individual	Genotype Loc1	Genotype Loc2	Trait
1	C/T	A/A	red
2	C/C	A/T	pink
3	C/T	T/T	white
4	T/T	T/T	white
5	C/C	A/T	pink
6	C/T	A/T	pink
7	T/T	A/T	pink
8	C/T	A/A	red


Loc1
Trait.Value
0
2
4
6
8
10
0.0
0.5
1.0
1.5
2.0
Number of 'A' alleles


Loc2
trait value
0
2
4
6
8
10
0.0
0.5
1.0
1.5
2.0
Number of 'A' alleles

Common Mapping Approaches:

- **Marker associations**
- QTL map
- RILs
- Hybrid mapping
- **BSA (bulk segregate analysis)/ poolseq**
- GWAS
- Fst outliers



Intro to mapping:

Background:

- *Mimulus* is a genus of wildflowers. Within the species *M. guttatus*, there exists 2 life histories: annuals and perennials.
- Genetic mapping has shown a large genomic region which differentiates annual and perennial forms of *M. guttatus* and explains about 40% of trait variation among these ecotypes
- Is this marker still associated with phenotypes in another, more distantly related species?

Intro to mapping:

The data:

- F2s between annual *M. guttatus* and perennial *M. decorus*
- Phenotyped for several life history traits, genotyped for their allele at the inversion.

Indiv	Inversion Ori	days.til.flowe	node of first	corolla lengtl	tube length	corolla width	stem circumf	first true leaf	first true leaf	internode ler	internode ler	# of stolons
12	AA	30	4	26.18	17.43	28.86	2.56	22.19	17.99	17.15	0.5	10
20	AA	27	4	23.57	13.14	19.84	1.13	8.18	10.1	5.28	3.23	4
21	AA	32	5	31.54	16.29	26.6	1.72	10.77	10.96	0.5	0.5	12
29	AA	23	2	26.2	14.74	20.86	0.86	13.85	14.05	11.1	8.21	0
30	AA	26	4	29.21	14.68	26.87	4.87	24.58	18.22	8.5	4.28	8
34	AA	30	5	29.36	14.48	26.98	3.23	29.91	24.34	6.52	0.5	8
37	AA	27	4	31.42	16.76	21.65	1.62	13.61	15.75	9.1	4.71	12
39	AA	29	5	31.16	16.58	28.82	2.21	27.18	22.16	16.08	0.5	10
43	AA	30	5	36.07	17.86	28.56	2.67	17.14	19.79	9.59	0.5	10
49	AA	27	4	36.93	17.9	28.79	2.3	30.05	22.17	14.97	2	6

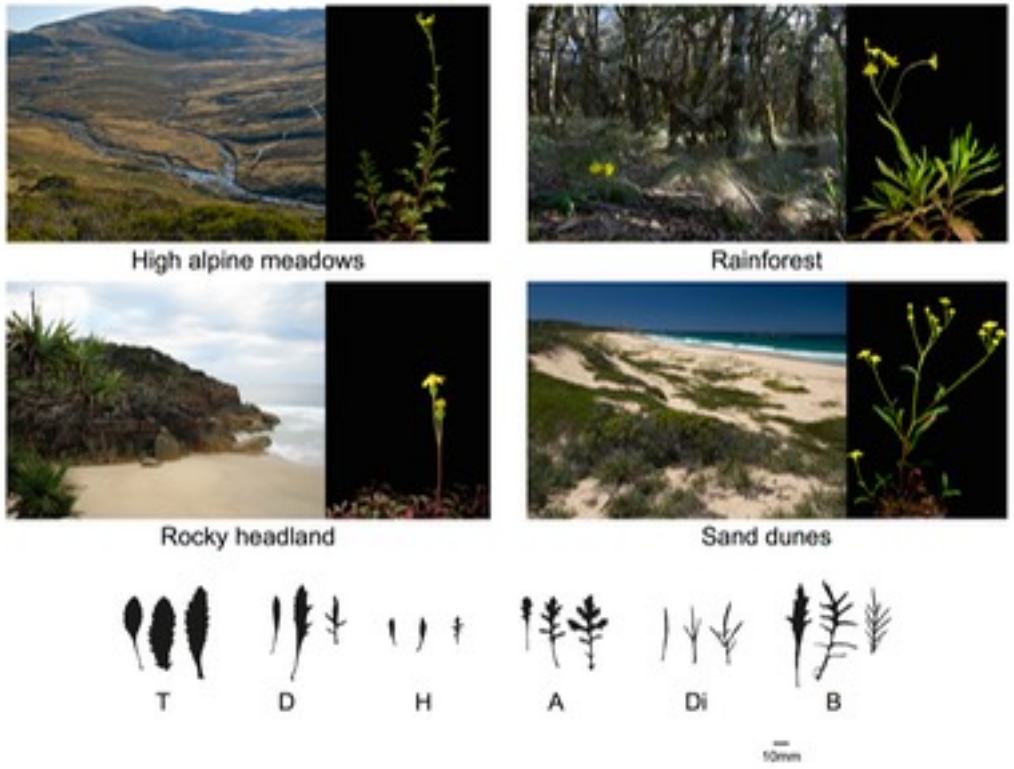
Use ggplot2 to make a figure to show the following:

1. Association between genotype and phenotype (choose 2 traits)

2. Trait correlation between phenotypes chosen in 1.

Part II: applying data viz rules to whole genome data!

- *Senecio lautus* is a wildflower that grows in Australia, inhabits many extremely different environments
- Different ecotypes are locally adapted to their home environment
- What is the genetic basis of these phenotype differences? What are the the loci involved in the ability to survive?
- The authors performed a **BSA**



The dataset:

- Publically available dataset from Roda *et al.* 2017
- .txt file where each row is a position in the genome
  - Position: the combination of LinkageGroup + Locus_Position
  - Comparison: Which of the three independent populations is being compared
  - Maa_pop1 andmaa_pop2: Major alleles found in Dune and Headlands environment
  - Counts: data one could use to calculate allele frequency
  - Fst: a measure of the difference between pools in allele frequency
  - P-value/ outlier_FET: other statistical measures of difference between pools

LocusID	Length	Coverage	LinkageGrou	Locus_Positi	Comparison	maa_pop1	maa_pop2	Count_maa_	Count_Total_	Count_maa_	Count_Total_	Fst	P-value_FET	Outlier_FET_	PositionMeth	LinkageGroup_Position
1000527	4893	9.05886	GR11	689	F8-C-S-Dune_	T	C	134	135	42	42	0.99253731	4.64E-40	1	noisy	44.832
1000527	4893	9.05886	GR11	641	F8-C-S-Dune_	G	A	134	135	42	42	0.99253731	4.64E-40	1	noisy	44.832
1000527	4893	9.05886	GR11	606	F8-C-S-Dune_	A	C	134	135	33	42	0.76982017	1.96E-26	0	noisy	44.832
1000527	4893	9.05886	GR11	751	F8-A-S-Dune_	G	A	45	76	34	34	0.58666667	1.19E-10	0	noisy	44.832
1000527	4893	9.05886	GR11	751	F8-C-S-Dune_	A	A	106	106	92	106	0.12380952	3.86E-05	0	noisy	44.832
1000527	4893	9.05886	GR11	751	F8-B-S-Dune_	G	G	26	47	81	106	0.07928435	0.0053824	0	noisy	44.832
1000527	4893	9.05886	GR11	771	F8-B-S-Dune_	A	A	47	47	101	106	0.03809524	0.15494459	0	noisy	44.832
1001169	1090	11.819266	GR20	776	F8-C-S-Dune_	T	G	30	30	16	26	0.6	1.28E-07	0	noisy	44.745
1001169	1090	11.819266	GR20	775	F8-C-S-Dune_	C	T	30	30	16	26	0.6	1.28E-07	0	noisy	44.745