Scientific Figure Design

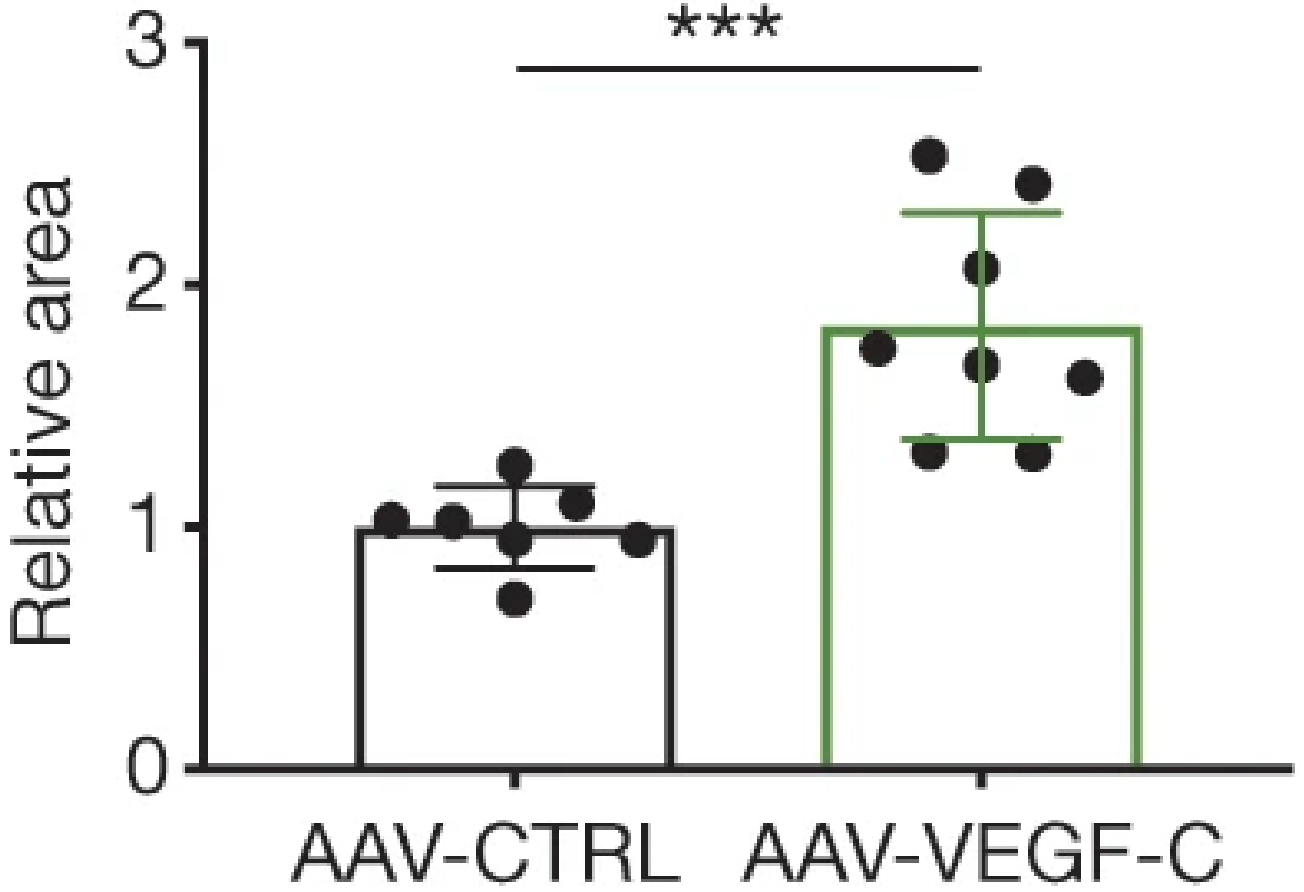
v2023-02

Simon Andrews, Anne Segonds-Pichon, Boo Virk, Jo Montgomery
simon.andrews@babraham.ac.uk
jo.montgomery@babraham.ac.uk

Babraham Bioinformatics

Figures are the way your science is presented to an audience

AAV-CTRL	AAV-VEGF-C
0.8245	1.3232
1.0136	2.5644
1.3224	1.4899
1.0128	1.512
0.9644	2.6002
0.9668	2.1132
1.2296	1.3228
1.0532	1.7566



What this course covers…

- Theory of data visualisation
  - Why do some figures work better than others?
  - Applying theory to common plot types
- Ethics of data representation
- Using graphic design

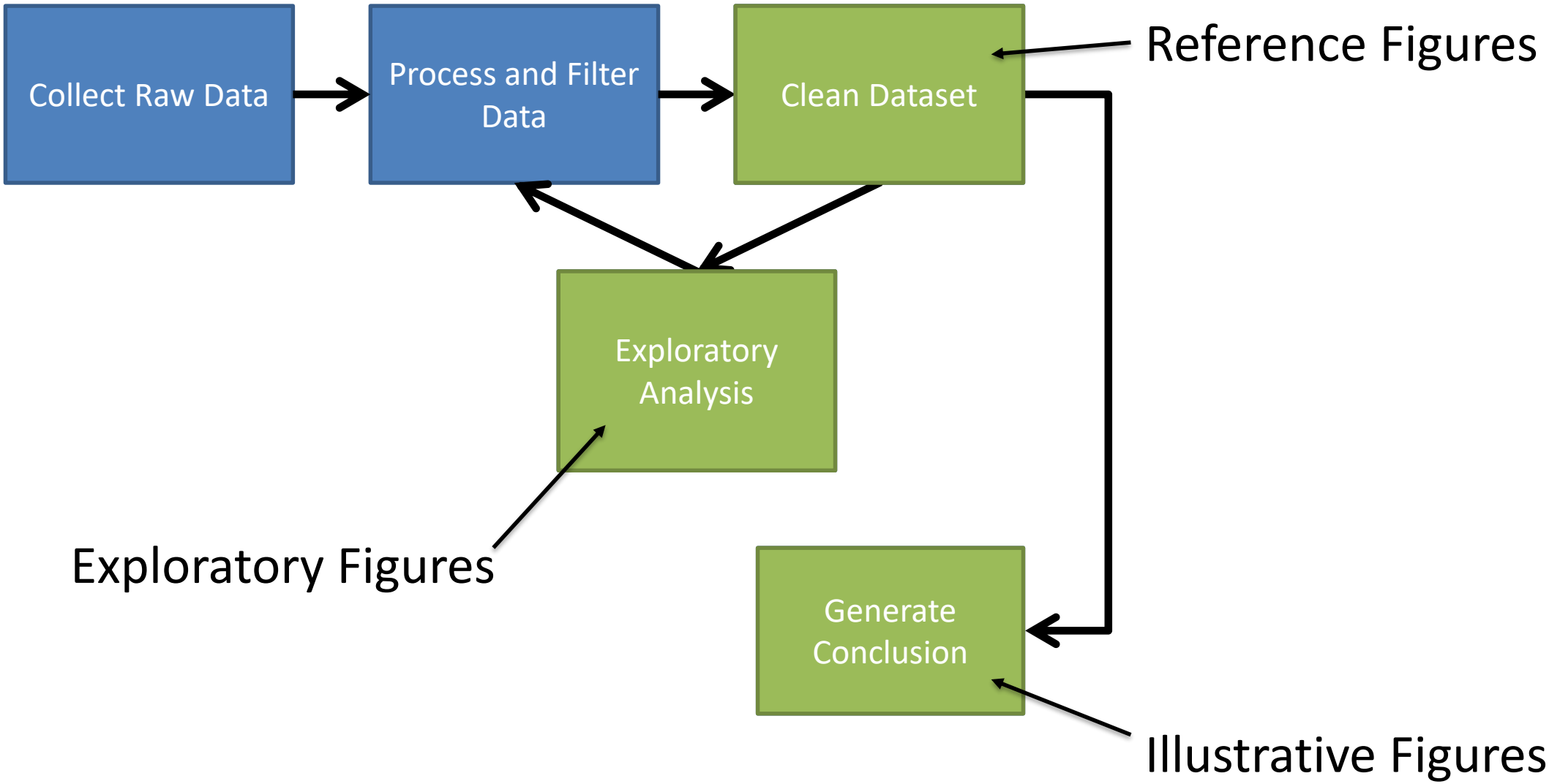
- Practical figure editing and compositing in Inkscape

What this course doesn't cover…

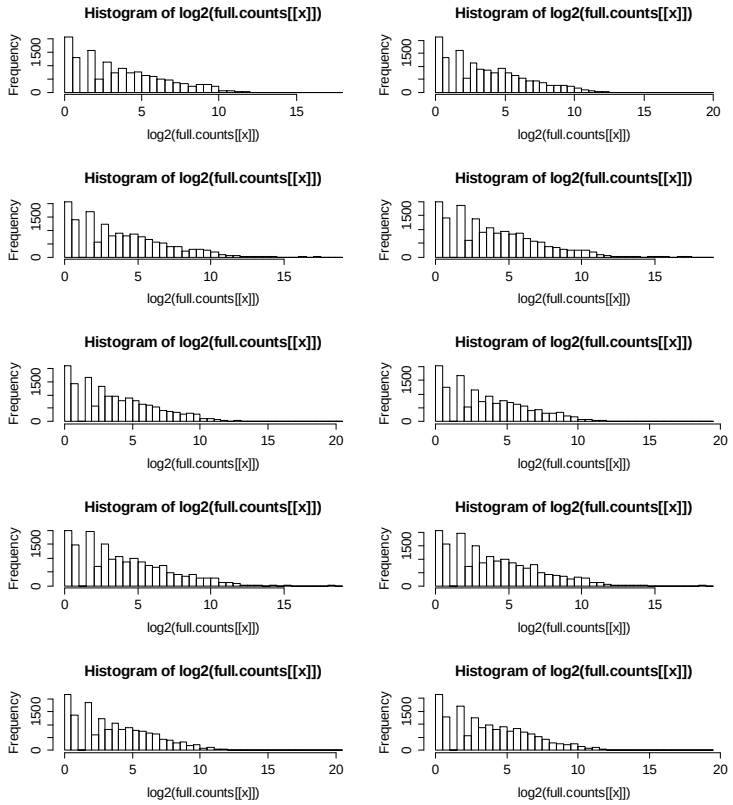
- How to draw graphs in specific programs

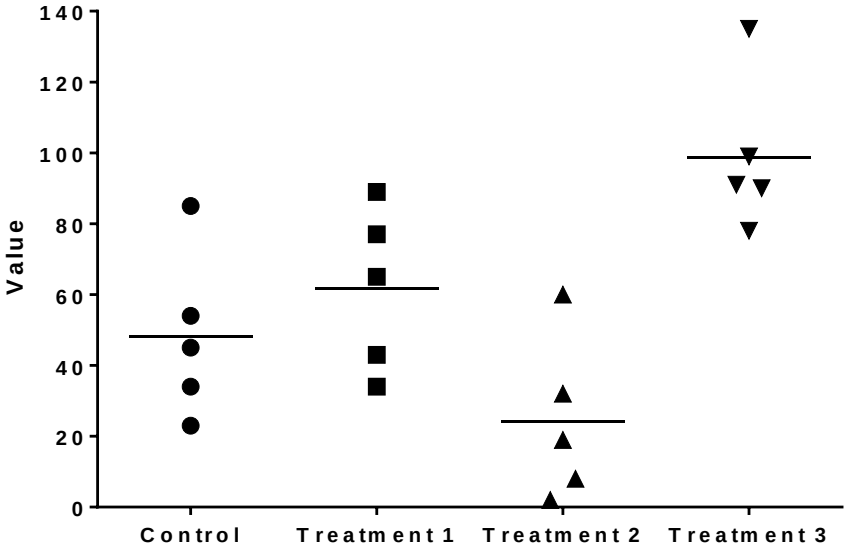
http://www.bioinformatics.babraham.ac.uk/training.html

Consider the requirements for a figure

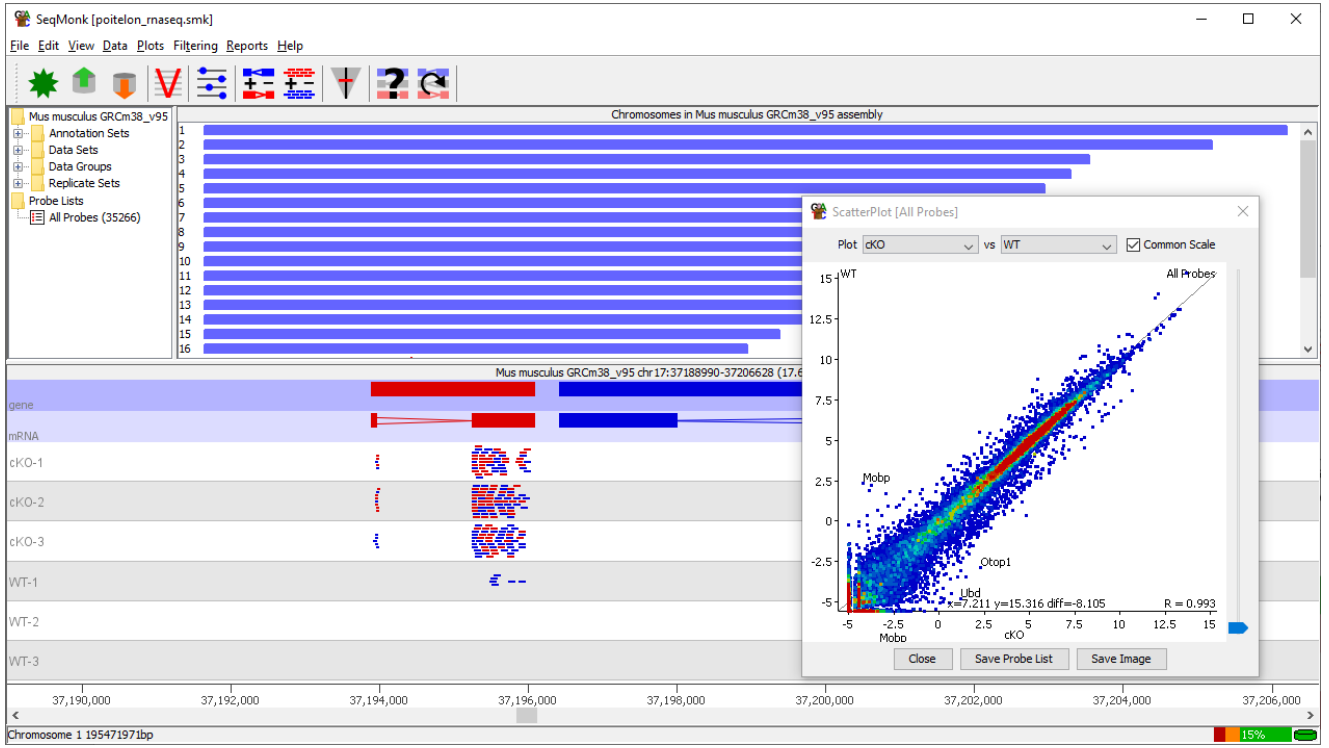


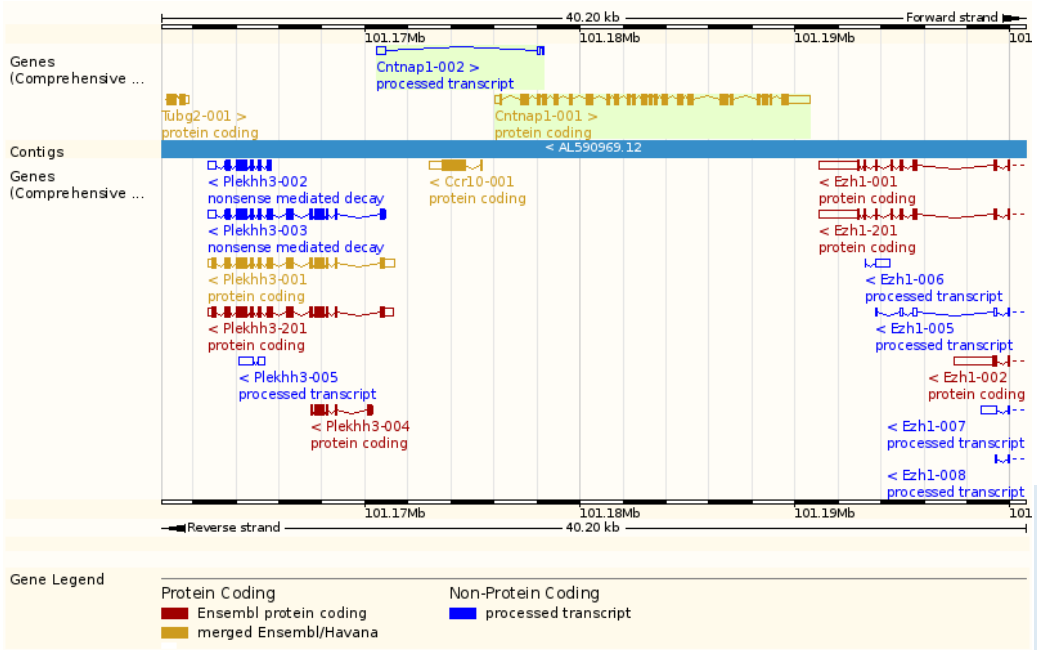
Exploratory figures





- Quick!
- Complete
- Interactive





Reference figures

- Complete
- Flexible

Feature Report

Chr	Start	End	Average ...	Feature	ID	Description	No. Su...	No. Pro...	Mean ...	StDev ...	Mea...	StD...
1	91285982	91726453		Centg2	ENSMUSG0000...	centaurin, gamma 2 [Source:M...	1	148	19.216	22.845	16.235	7.427
1	91759133	91762347		Gbx2	ENSMUSG0000...	gastrulation brain homeobox 2...	1	2	321.5	137.886	89.343	36.373
1	91783849	91845748		Asb18	ENSMUSG0000...	ankyrin repeat and SOCS box-...	1	22	22.227	30.095	47.379	39.908
1	91873797	91920700		Iqca	ENSMUSG0000...	IQ motif containing with AAA d...	1	17	11.706	5.665	21.977	6.709
1	91958628	91976324		Iqca	ENSMUSG0000...	IQ motif containing with AAA d...	1	7	7.429	3.207	23.786	4.68
1	92035174	92047459		Cxcr7	ENSMUSG0000...	chemokine (C-X-C motif) recep...	1	5	30.8	31.523	20.034	6.862
1	92433385	92443499		Cops8	ENSMUSG0000...	COP9 (constitutive photomorp...	1	5	47.4	70.692	13.266	3.371
1	92505565	92510197		EG433332	ENSMUSG0000...	predicted gene, EG433332 [So...	1	2	7.5	3.536	20.305	1.914
1	92596517	92674334		Col6a3	ENSMUSG0000...	procollagen, type VI, alpha 3 [...	1	27	8.296	3.821	26.773	8.431
1	92745509	92781547		Mlph	ENSMUSG0000...	melanophilin [Source:MarkerSy...	1	13	17.077	13.175	32.384	13.865
1	92783511	92784418		LOC633040	ENSMUSG0000...	No description	1	2	14	1.414	28.427	9.572
1	92788541	92800026		Rab17	ENSMUSG0000...	RAB17, member RAS oncogen...	1	5	19.6	15.388	40.881	12.892
1	92829210	92958602		Lrrfip1	ENSMUSG0000...	leucine rich repeat (in FLII) int...	1	45	23.156	40.155	21.93	21.654
1	92975509	93001201		Gm817	ENSMUSG0000...	gene model 817, (NCBI) [Sour...	1	10	12.8	14.65	14.484	7.221
1	93010229	93054126		Ramp1	ENSMUSG0000...	receptor (calcitonin) activity m...	1	15	15.067	13.593	39.979	20.337
1	93054380	93055601		9130218E1...	ENSMUSG0000...	RIKEN cDNA 9130218E19 gen...	1	2	12.5	2.121	35.872	0.957
1	93114058	93115422		Ube2f	ENSMUSG0000...	ubiquitin-conjugating enzyme ...	1	2	5	4.243	5.415	3.829
1	93126451	93127342		ENSMUSG0...	ENSMUSG0000...	No description	1	1	218	0	14.89	0
1	93128744	93151480		Scly	ENSMUSG0000...	selenocysteine lyase [Source:...	1	9	37.889	68.054	15.492	7.038
1	93152481	93178709		Espnl	ENSMUSG0000...	espin-like [Source:MarkerSymb...	1	10	20.8	28.44	66.33	64.176
1	93181479	93192810		Klhl30	ENSMUSG0000...	kelch-like 30 (Drosophila) [Sou...	1	5	11.6	5.177	41.423	5.029
1	93196836	93204622		BC056923	ENSMUSG0000...	cDNA sequence BC056923 [So...	1	4	61.75	84.626	27.412	22.199
1	93206237	93229189		Ilkap	ENSMUSG0000...	integrin-linked kinase-associat...	1	9	42.222	66.944	11.431	2.451
1	93241889	93243628		Hes6	ENSMUSG0000...	hairy and enhancer of split 6 (...	1	2	190	31.113	71.745	26.801
1	93246388	93289702		Per2	ENSMUSG0000...	period homolog 2 (Drosophila) ...	1	16	33.562	60.427	38.834	26.818
1	93325073	93358670		Traf3ip1	ENSMUSG0000...	TNF receptor-associated facto...	1	12	30.083	72.455	13.086	6.513
1	93371120	93389585		Asb1	ENSMUSG0000...	ankyrin repeat and SOCS box-...	1	7	51.857	105.506	12.957	5.841
1	93631867	93678434		Twist2	ENSMUSG0000...	twist homolog 2 (Drosophila) [...	1	17	18.765	31.888	53.51	76.591
1	93760418	93760711		Hdac4	ENSMUSG0000...	histone deacetylase 4 [Source...	1	1	28	0	13.537	0
1	93763364	93885379		Hdac4	ENSMUSG0000...	histone deacetylase 4 [Source...	1	41	12.024	4.396	12.183	4.227
1	94269894	94304205		Ndufa10	ENSMUSG0000...	NADH dehydrogenase (ubiquin...	1	12	41.75	110.694	12.747	3.288

Close Save

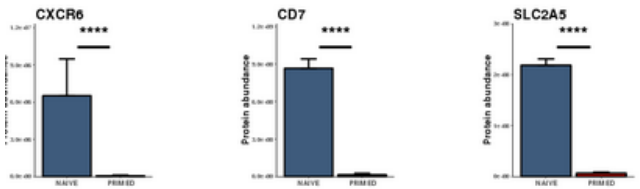
	Uniprot id	Gene	Protein names	Gene names	Pep Count	log2 fold change	q-value	GO terms
			All	All				All
1	O00574	CXCR6	C-X-C chemokine...	CXCR6 BONZO STR...	2	5.718	1.5e-7	G protein-coupled receptor sig...
2	P19397	CD53	Leukocyte surfa...	CD53 MOX44 TSPA...	3	5.408	4.83e-7	neutrophil degranulation [GO:0...
3	P09564	CD7	T-cell antigen ...	CD7	2	5.369	1.15e-9	T cell activation [GO:0042110]...
4	P05187	ALPP	Alkaline phosph...	ALPP PLAP	13	5.225	1.58e-9	0;anchored component of membra...
5	Q07075	ENPEP	Glutamyl aminop...	ENPEP	40	5.153	3.83e-8	angiogenesis [GO:0001525]; ang...
6	P22732	SLC2A5	Solute carrier ...	SLC2A5 GLUT5	9	5.081	3.26e-9	carbohydrate metabolic process...
7	P15509	CSF2RA	Granulocyte-mac...	CSF2RA CSF2R CS...	14	5.033	3.08e-9	MAPK cascade [GO:0000165]; cel...
8	P15328	FOLR1	Folate receptor...	FOLR1 FOLR	2	4.848	4.39e-9	COPII vesicle coating [GO:0048...
9	P13726	F3	Tissue factor (...	F3	8	4.836	4.47e-8	activation of blood coagulatio...
10	Q9BYF1	ACE2	Angiotensin-con...	ACE2 UNQ868/PRO...	4	4.83	3.93e-8	angiotensin maturation [GO:000...

Showing 1 to 10 of 1,715 entries

Previous 1 2 3 4 5 ... 172 Next

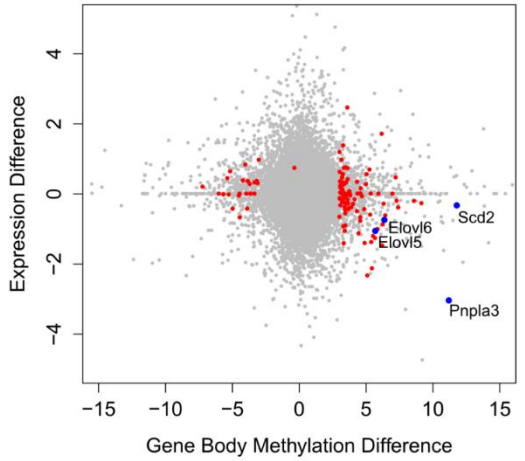
Download Table Clear Filters Clear selected rows

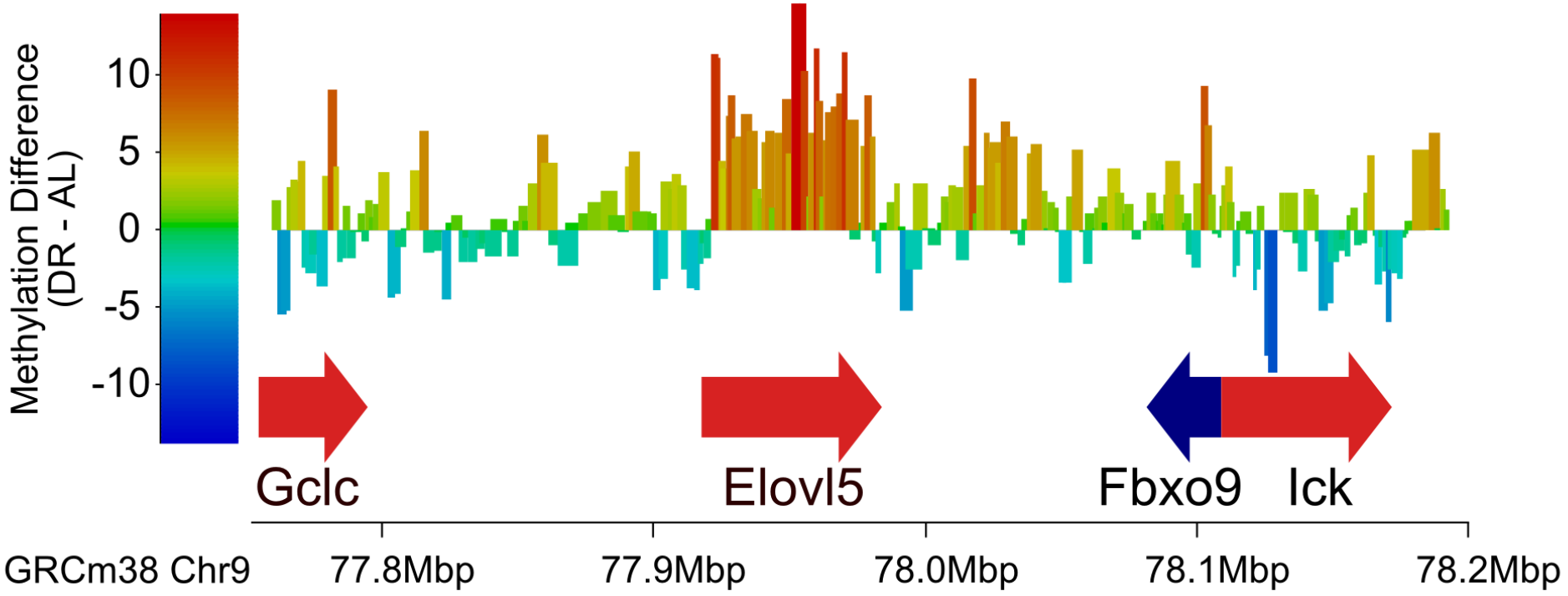
Select up to 6 rows in the table to display plots

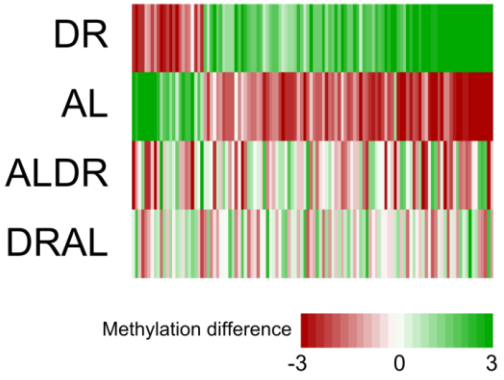


Illustrative figures

- Simple
- Clear
- Pretty







What makes a good figure?

- Has a clear purpose and message
  - Helps to tell a story
  - Adds to the text, and links to it
- Is focused
  - Don’t confuse one message with another
- Is easy to interpret correctly
  - Good data visualisation
  - Good design
- Is an honest and true reflection of the data

The theory of data visualisation

Simon Andrews, Phil Ewels

simon.andrews@babraham.ac.uk

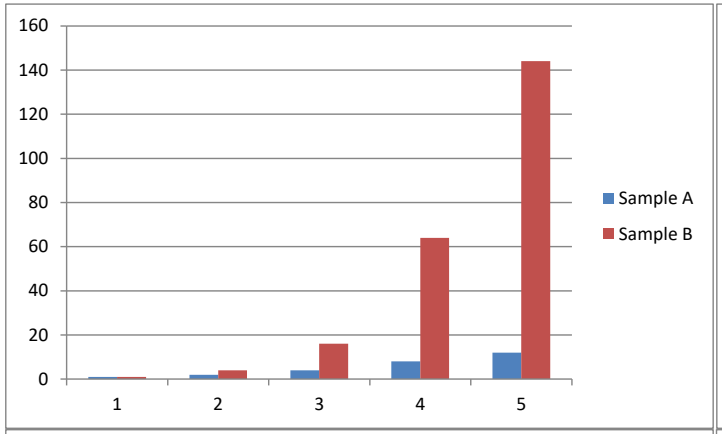
Babraham Bioinformatics

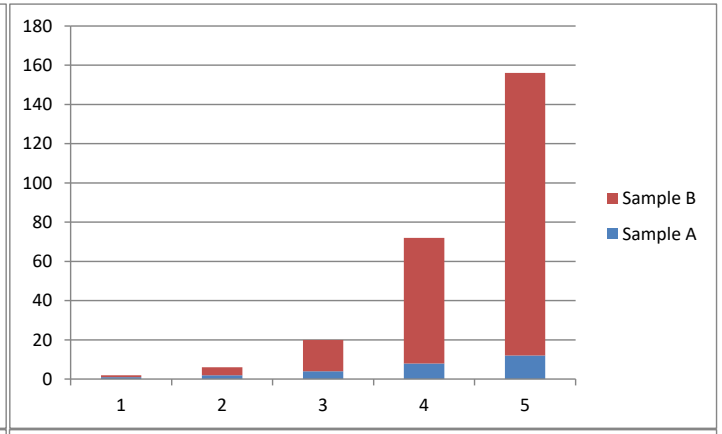
Data Visualisation

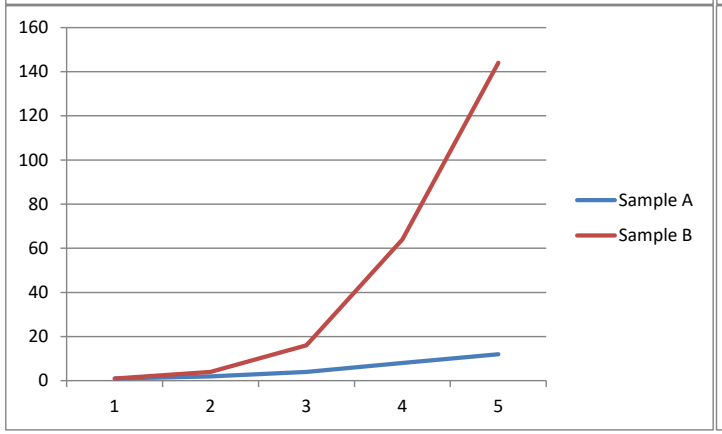
- A scientific discipline involving the creation and study of the visual representation of data whose goal is to communicate information clearly and efficiently to users.

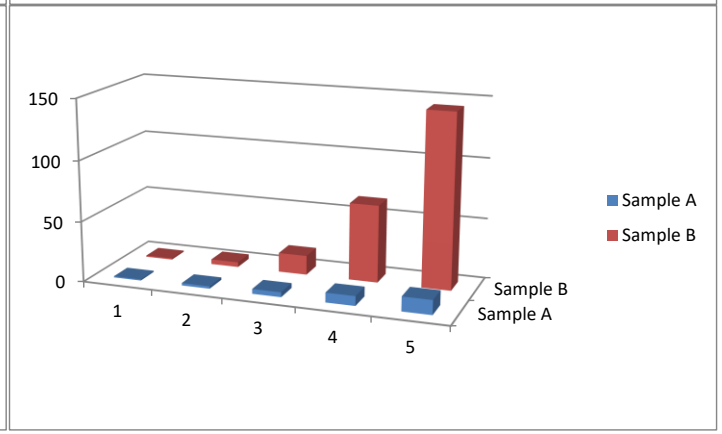
- Data Visualisation is both an art and a science.

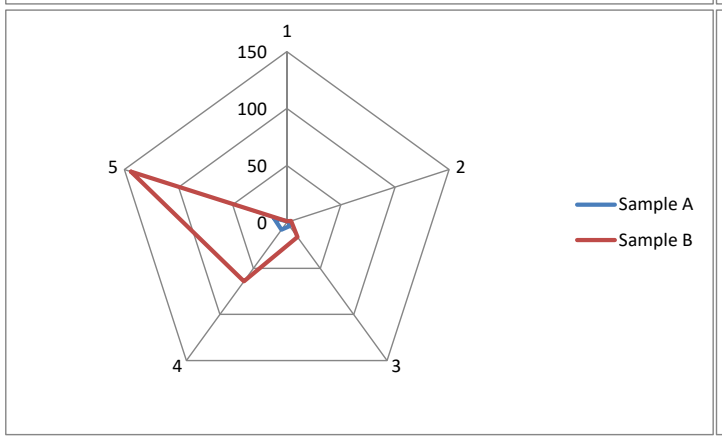
Sample A	Sample B
1	1
2	4
4	16
8	64
12	144

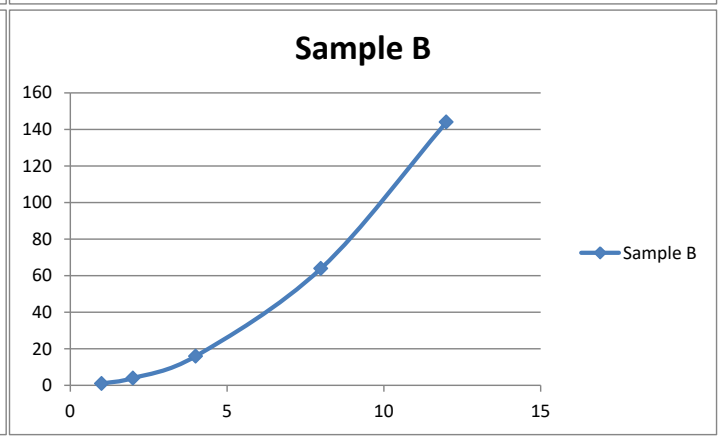


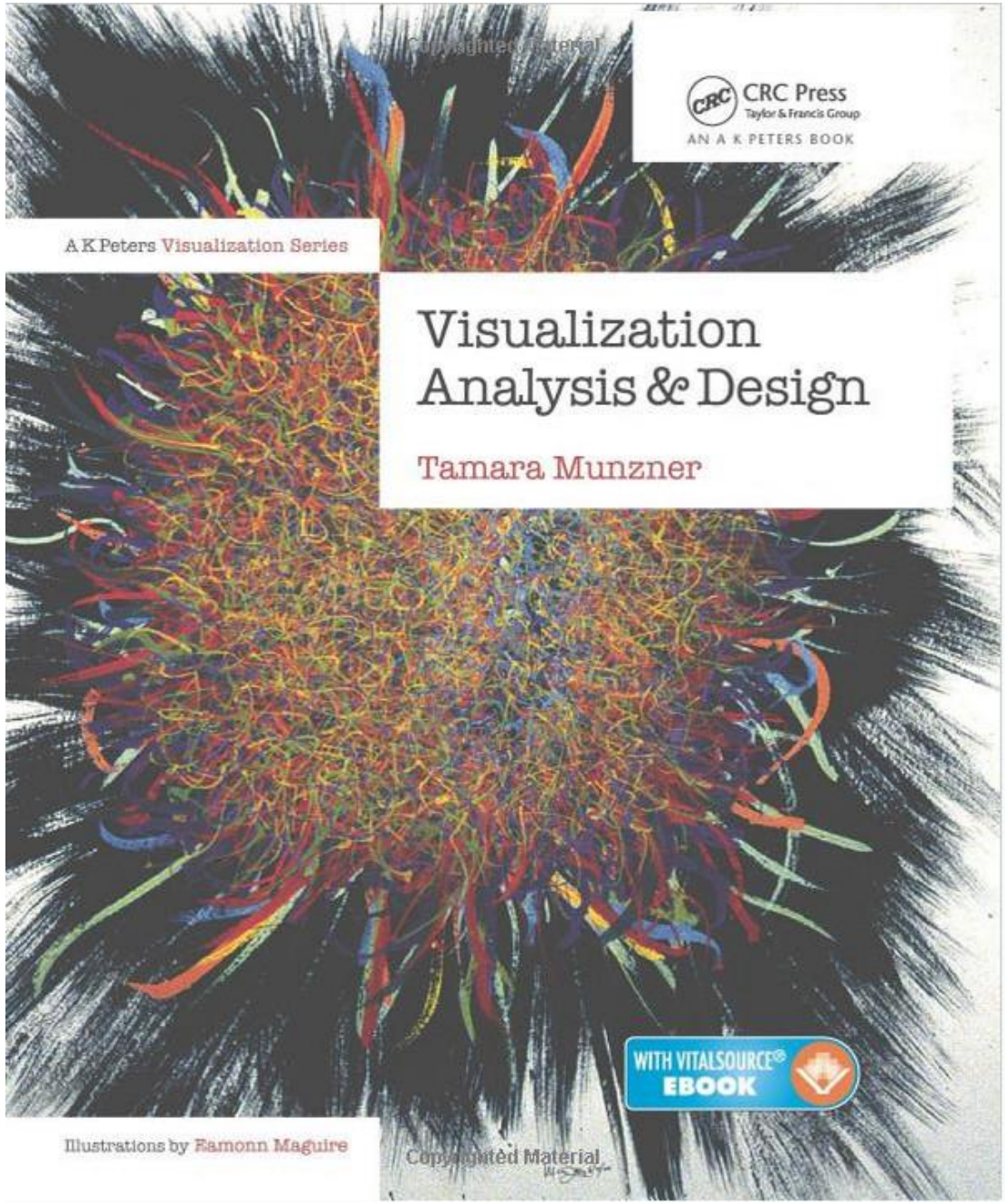






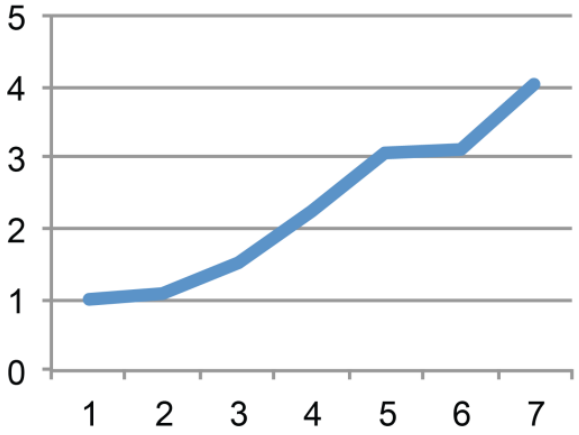




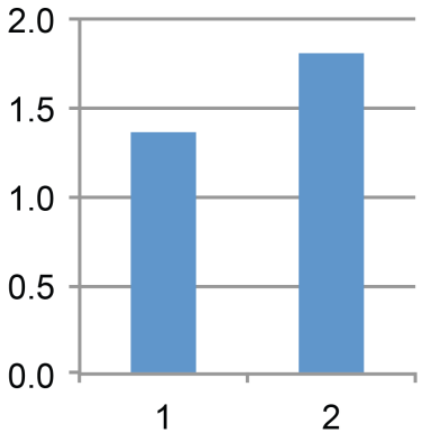


ISBN-10: 1466508914
http://www.cs.ubc.ca/~tmm/talks.html

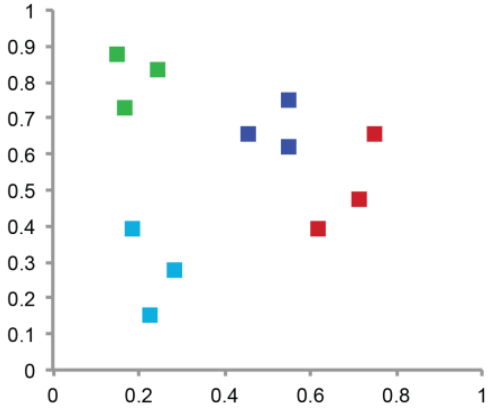
Different representations have common elements



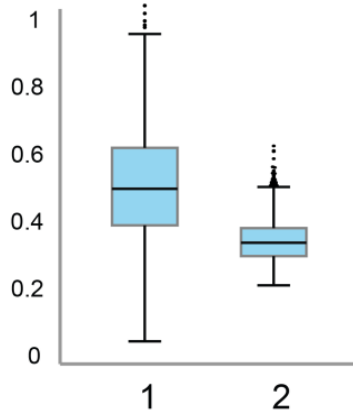
Relationship



Comparison



Composition

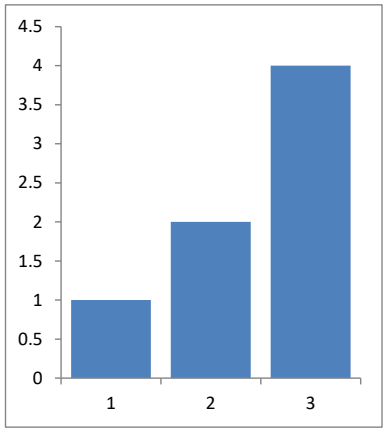


Distribution

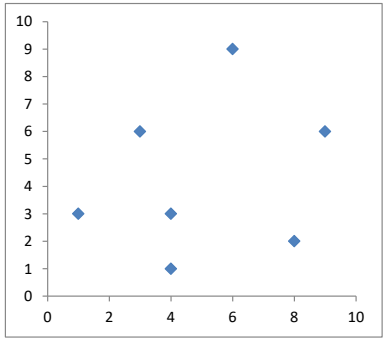
Marks and Channels

- Marks
  - Geometric primitives
    - Lines
    - Points
    - Areas
  - Used to represent data sets
- Channels
  - Graphical appearance of a mark
    - Colour
    - Length
    - Position
    - Angle
  - Used to encode data

Figures are a combination of marks and channels

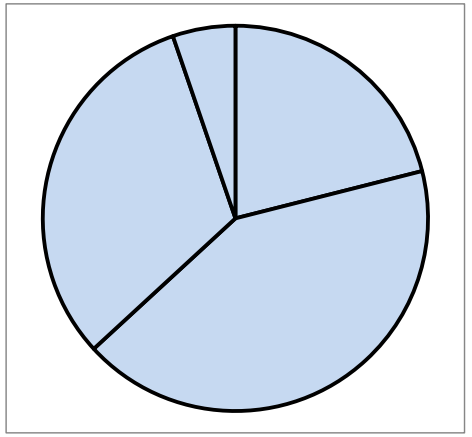


1 Mark = Rectangle
1 Channel = Length of longest side



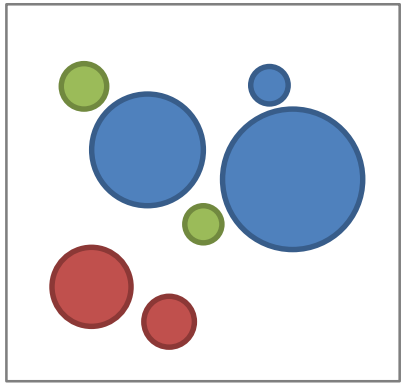
1 Mark = Diamond shape
2 Channels = X position, Y position

1 Mark = Circle segment
1 Channel = Angle



1 Mark = Circle
4 Channels:

X position
Y position
Area
Colour

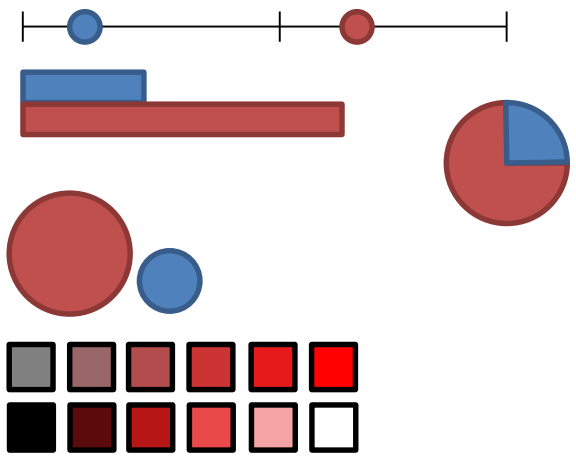


Golden Rules

- Expressiveness
  - Match the properties of the data and channel
- Effectiveness
  - Encode the most important information with the most effective channel

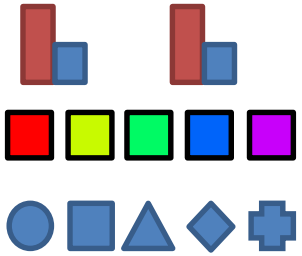
Types of channel

- Quantitative
  - Position on scale
  - Length
  - Angle
  - Area
  - Colour (saturation)
  - Colour (lightness)



- Quantitative
  - Weight
  - Length
  - Height
  - Expression
  - Time
  - Density

- Qualitative
  - Spatial Grouping
  - Colour (hue)
  - Shape

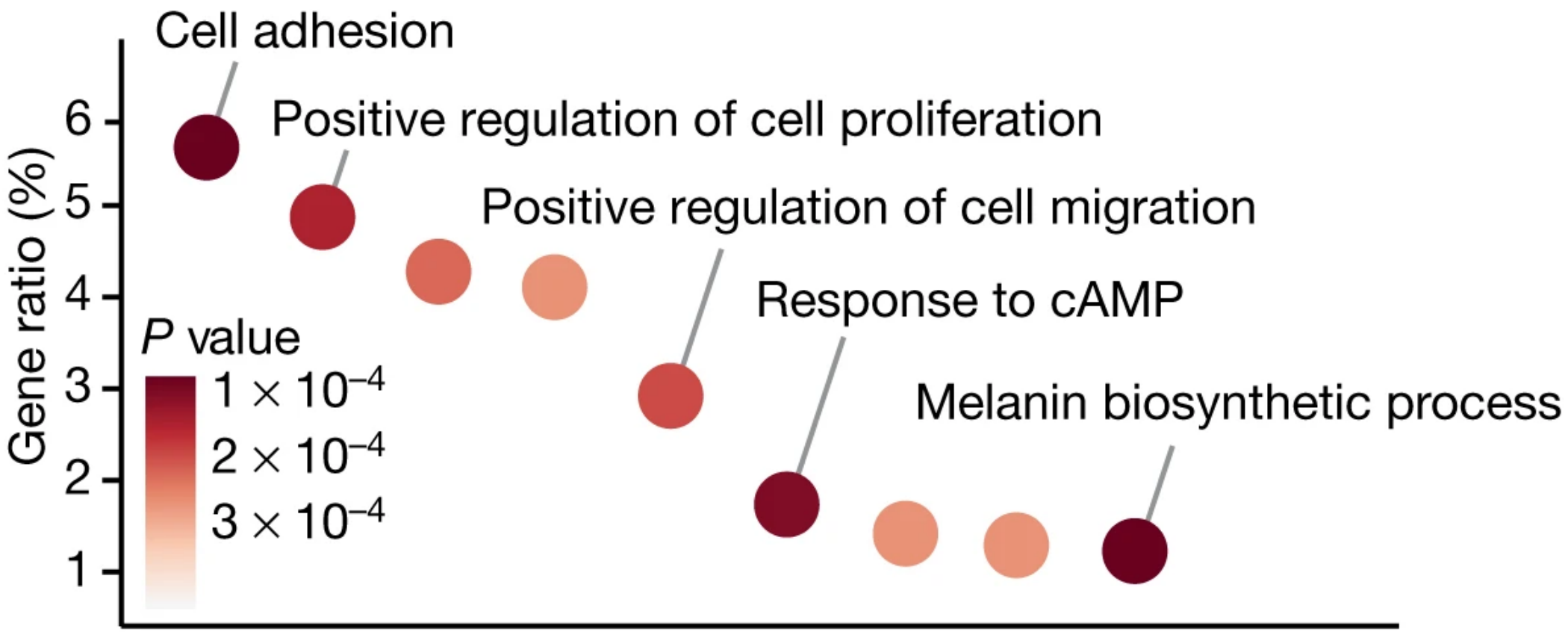


- Qualitative
  - Treatment
  - Genotype
  - Batch

Golden Rules

- Expressiveness
  - Match the properties of the data and channel
- Effectiveness
  - Encode the most important information with the most effective channel

Matching the data and channel



Colour

- Only channel to appear in both Qualitative and Quantitative
  - Technical representations of colour
    - Red + Green + Blue (RGB)
    - Cyan + Magenta + Yellow + Black (CMYK)
  - Perceptual representation of colour
    - Hue + Saturation + Lightness (HSL)

HSL Representation

- Hue = Shade of colour = Qualitative
- Saturation = Amount of colour = Quantitative
- Lightness = Amount of white = Quantitative

- Humans have no innate quantitative perception of hue but we have learned some (cold – hot, rainbow etc.)

- Our perception of hue is not linear

Types of channel

- Quantitative
  - Colour (saturation)
  - Colour (lightness)

- Qualitative
  - Colour (hue)

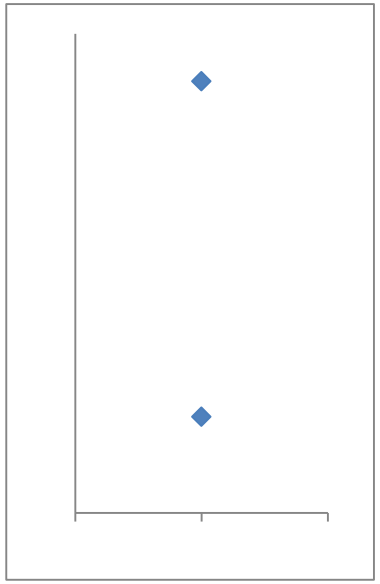
In a single plot you should modify only ONE colour parameter

Golden Rules

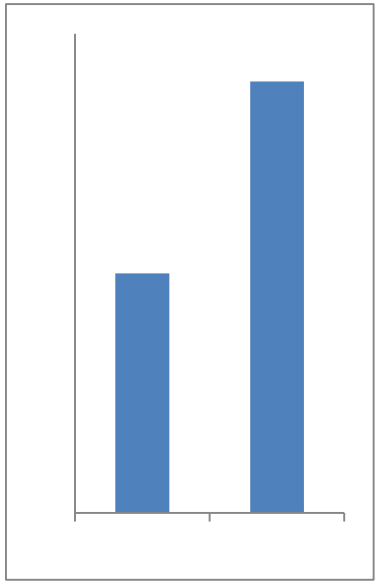
- Expressiveness
  - Match the properties of the data and channel

- Effectiveness
  - Encode the most important information with the most effective channel

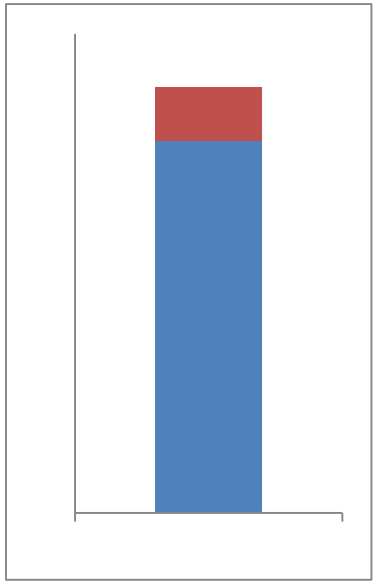
Effectiveness of quantitative channels



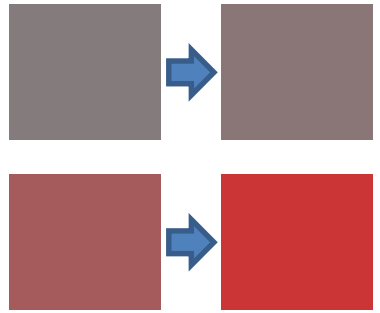
4.5X



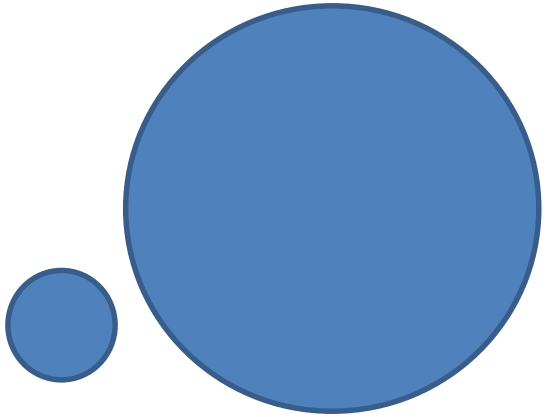
1.8X



7X

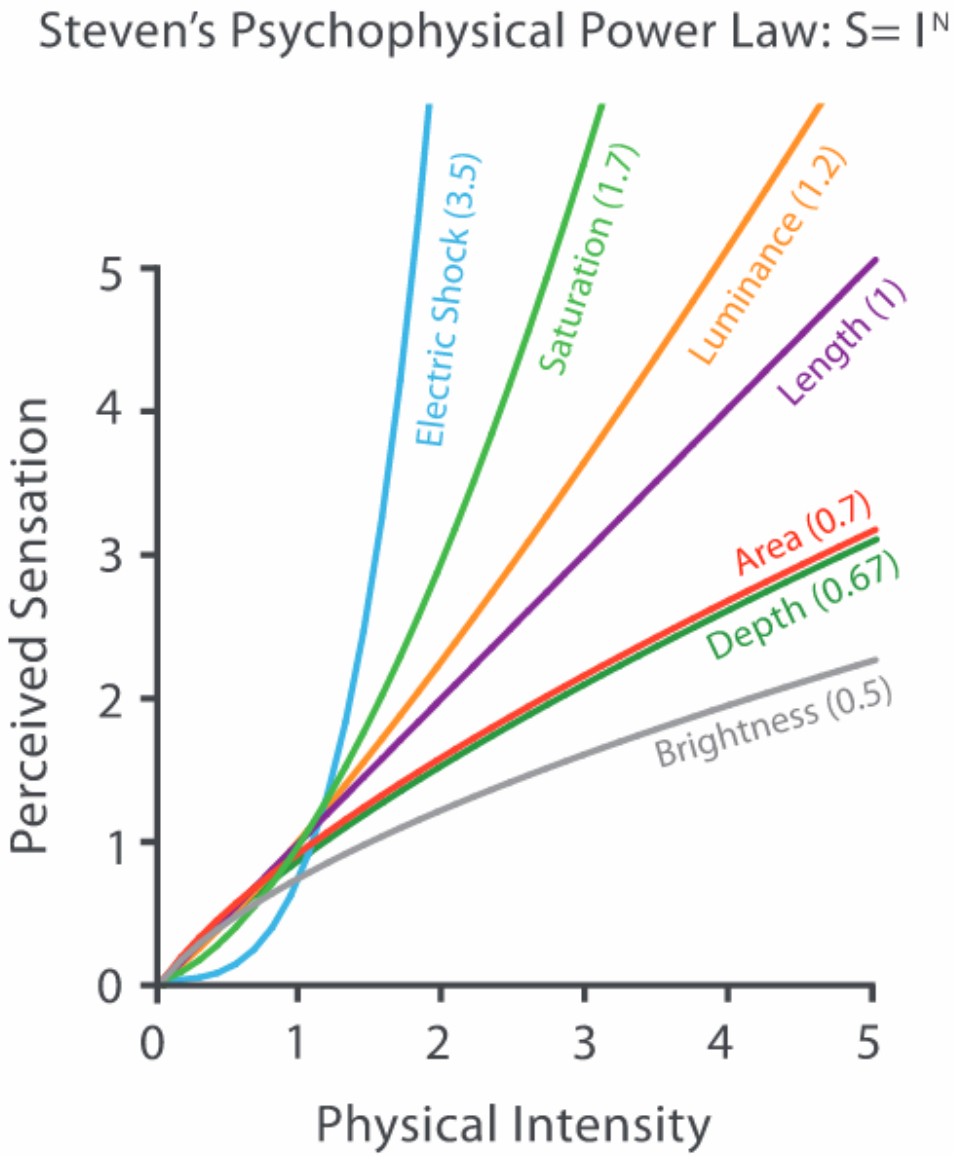


2X



16X

Quantitation Perception



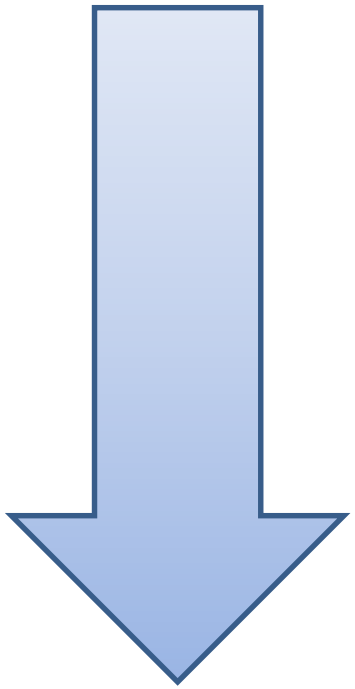
Golden Rules

- Effectiveness
  - Encode the most important information with the most effective channel

- Expressiveness
  - Match the properties of the data and channel

Most Quantitative Representations

Good quantitation

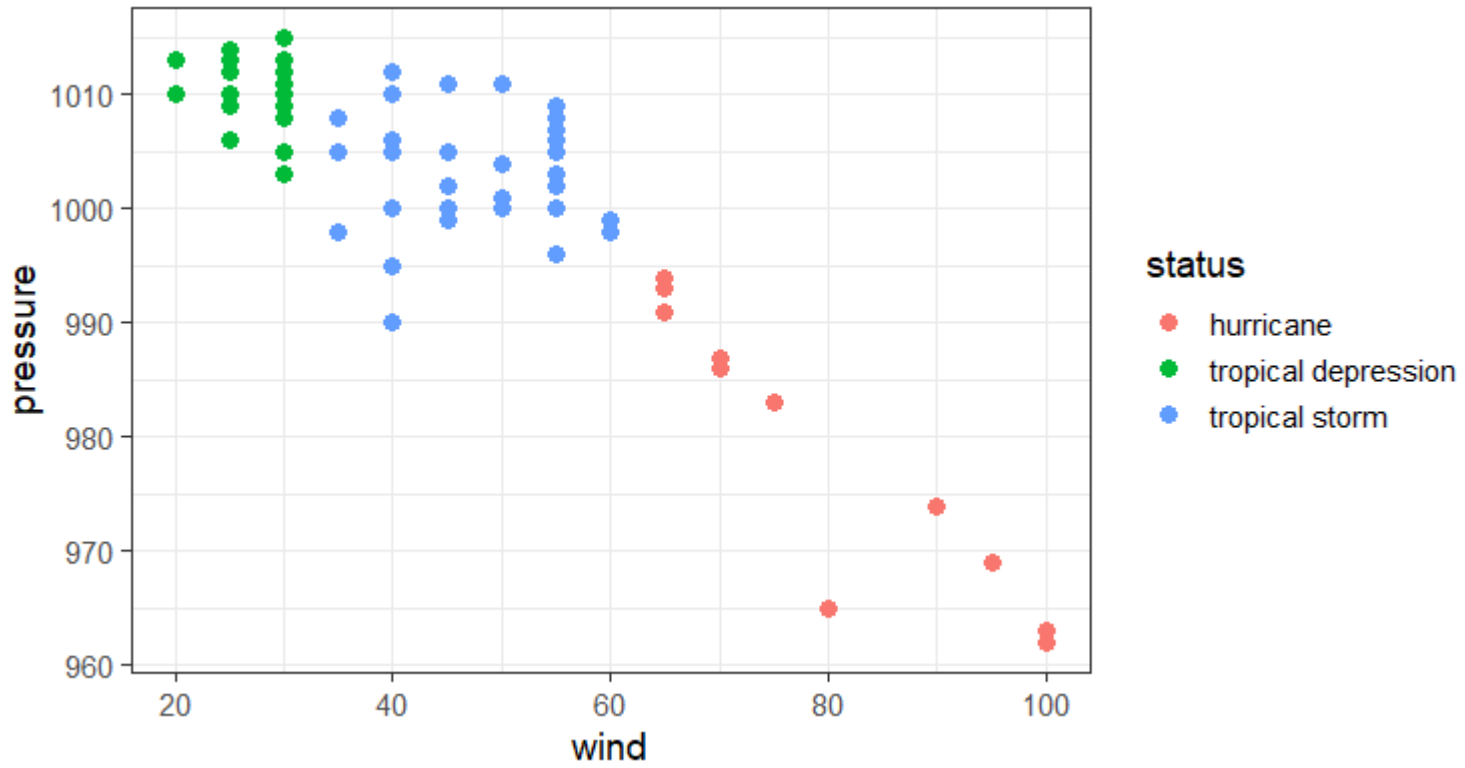


Poor quantitation

- Bar chart
- Stacked bar chart with common start
- Stacked bar chart with different starts
- Pie charts
- Bubble plots (circular area)
- Rectangular area
- Colour (luminance)
- Colour (saturation)

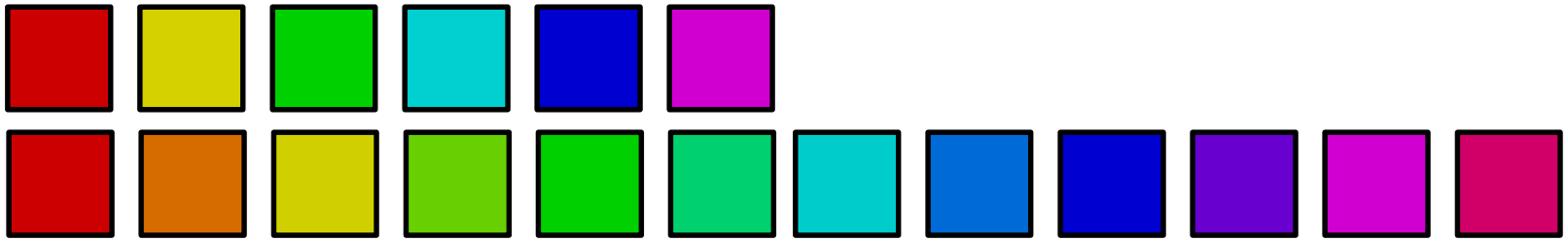
Effectiveness of Qualitative Channels

- If you encode categorical data are the differences between categories easy for the user to perceive correctly?



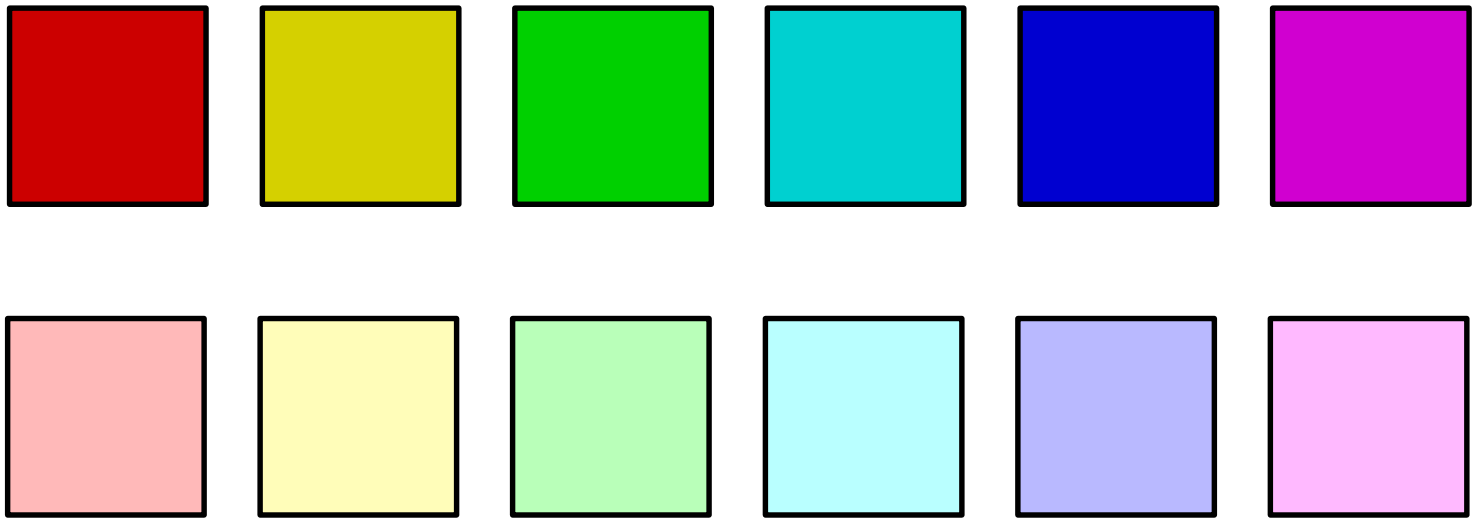
Colour Discrimination

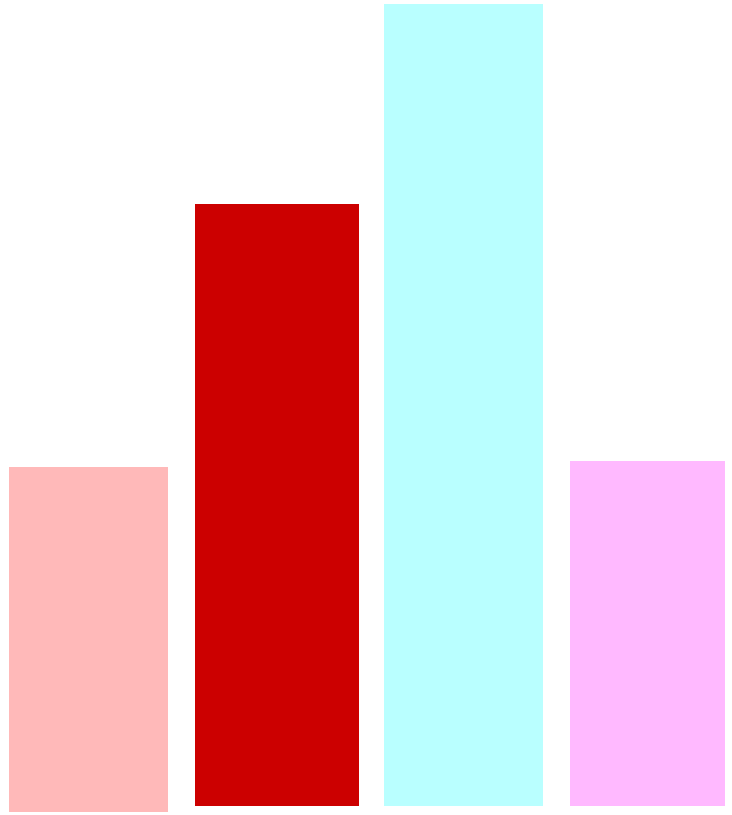
- How many colours can you discriminate?



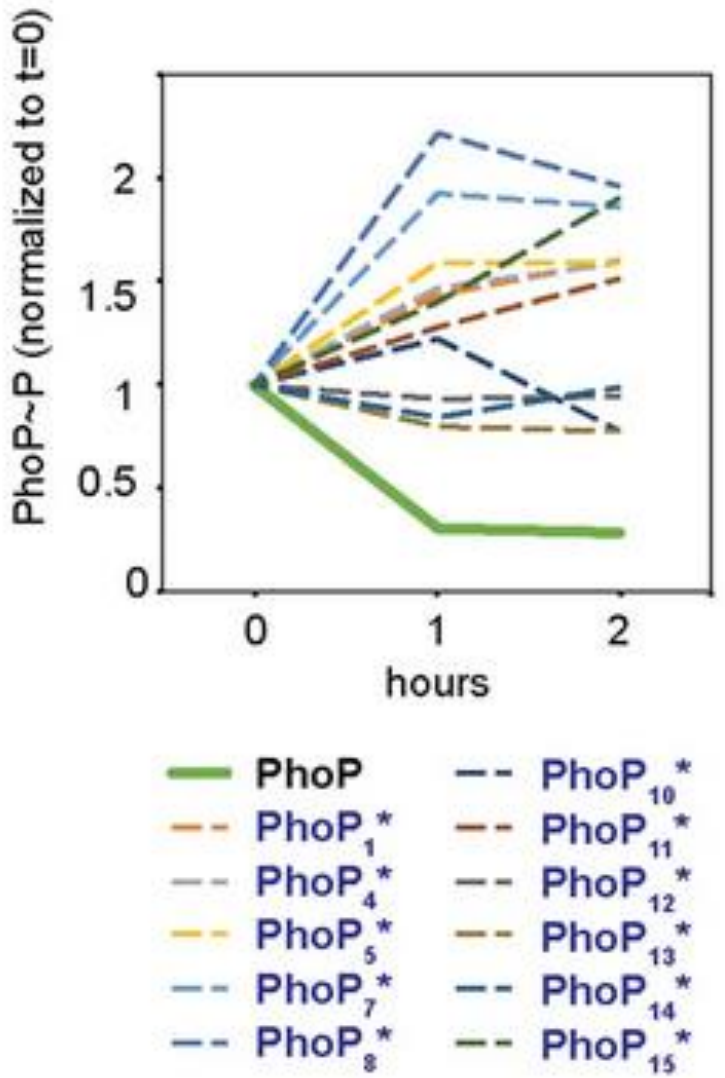
Colour Discrimination

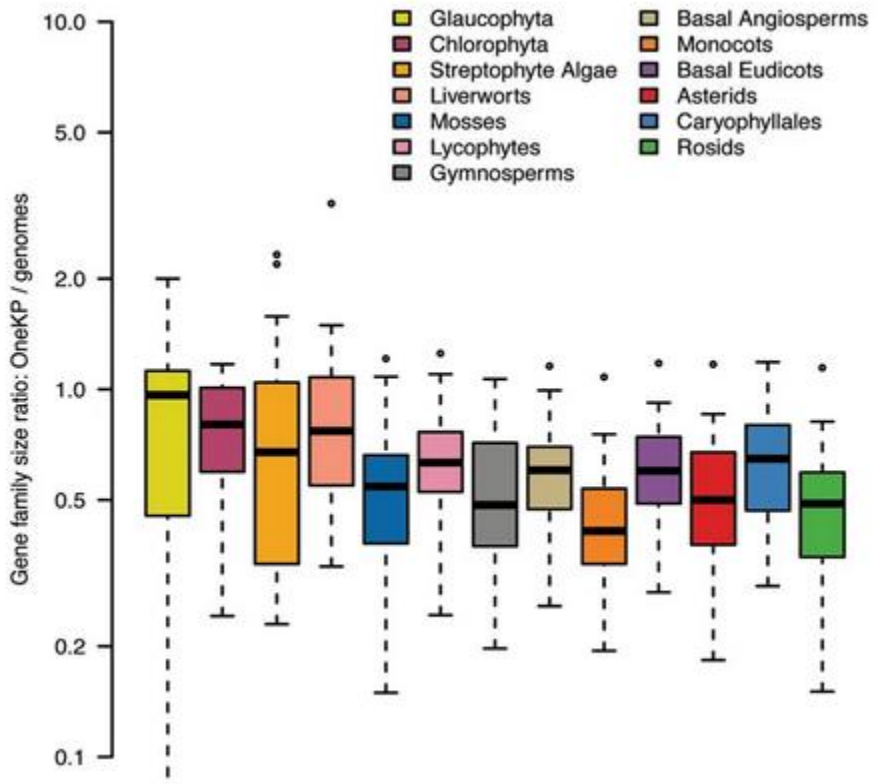
- How many colours can you discriminate?

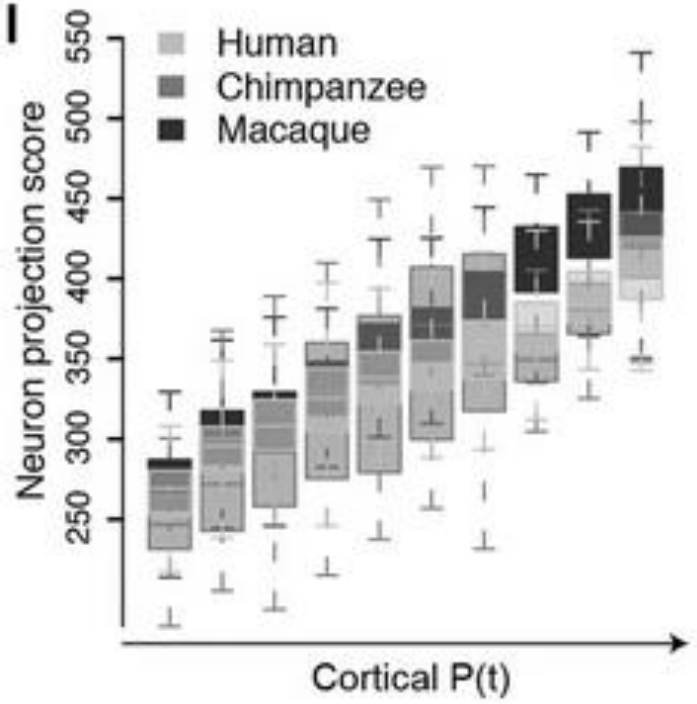




Colour Discrimination

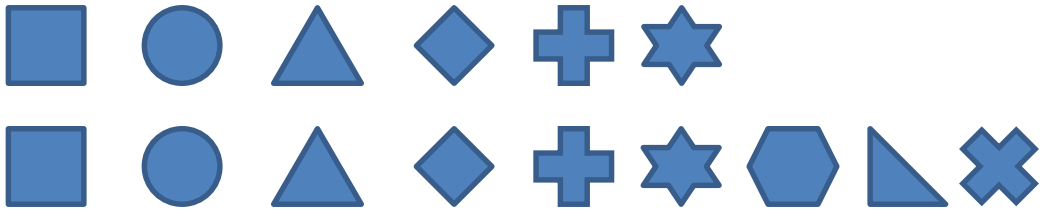






Qualitative Discrimination

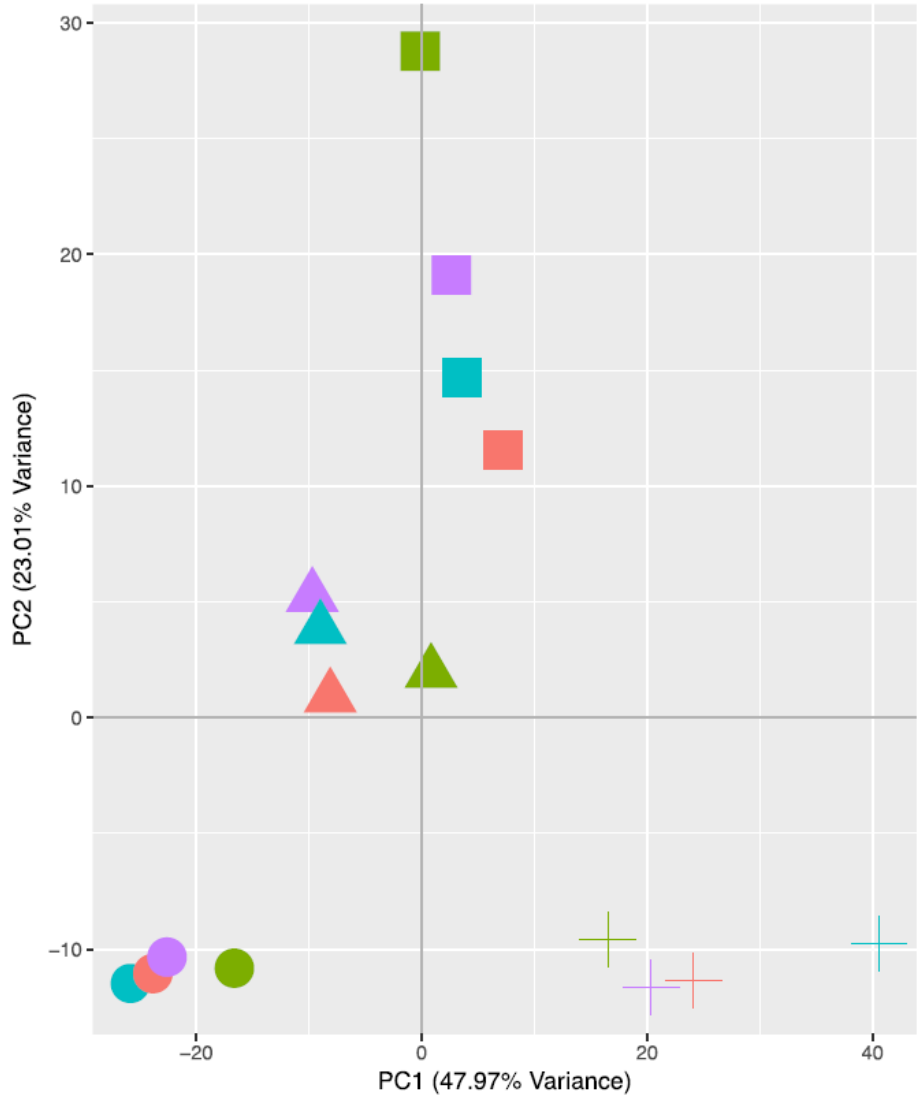
- How many (fillable) shapes can you discriminate?



- Can combine shape with colour, but you need to maintain similar fillable areas

Qualitative Discrimination

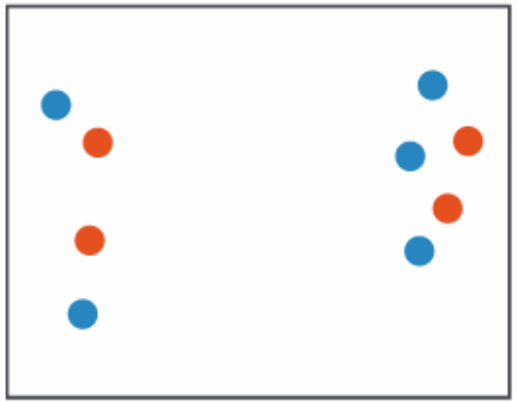
- You can combine shape with colour, but you need to maintain similar fillable areas



Separability

Adding channels can adversely affect the effectiveness of existing channels

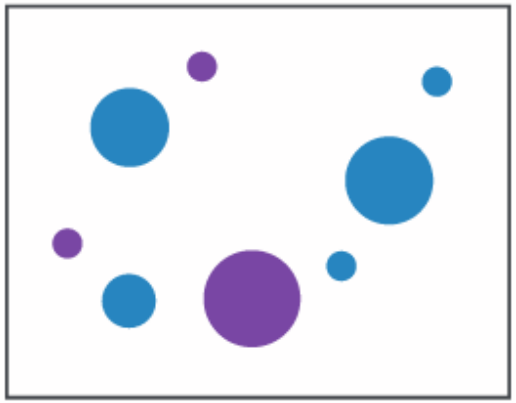
Position
+ Hue (Color)



Fully separable

There is no confusion between the two channels

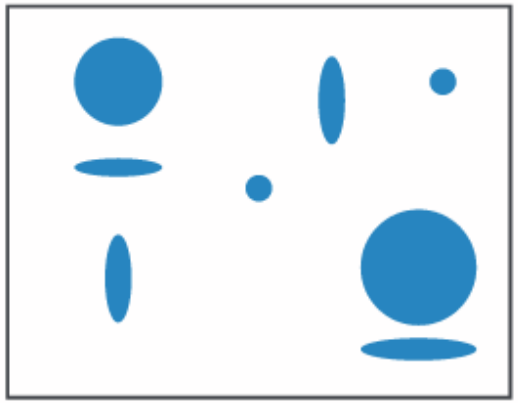
Size
+ Hue (Color)



Some interference

Larger points are easier to discriminate than smaller ones

Width
+ Height



Some/significant interference

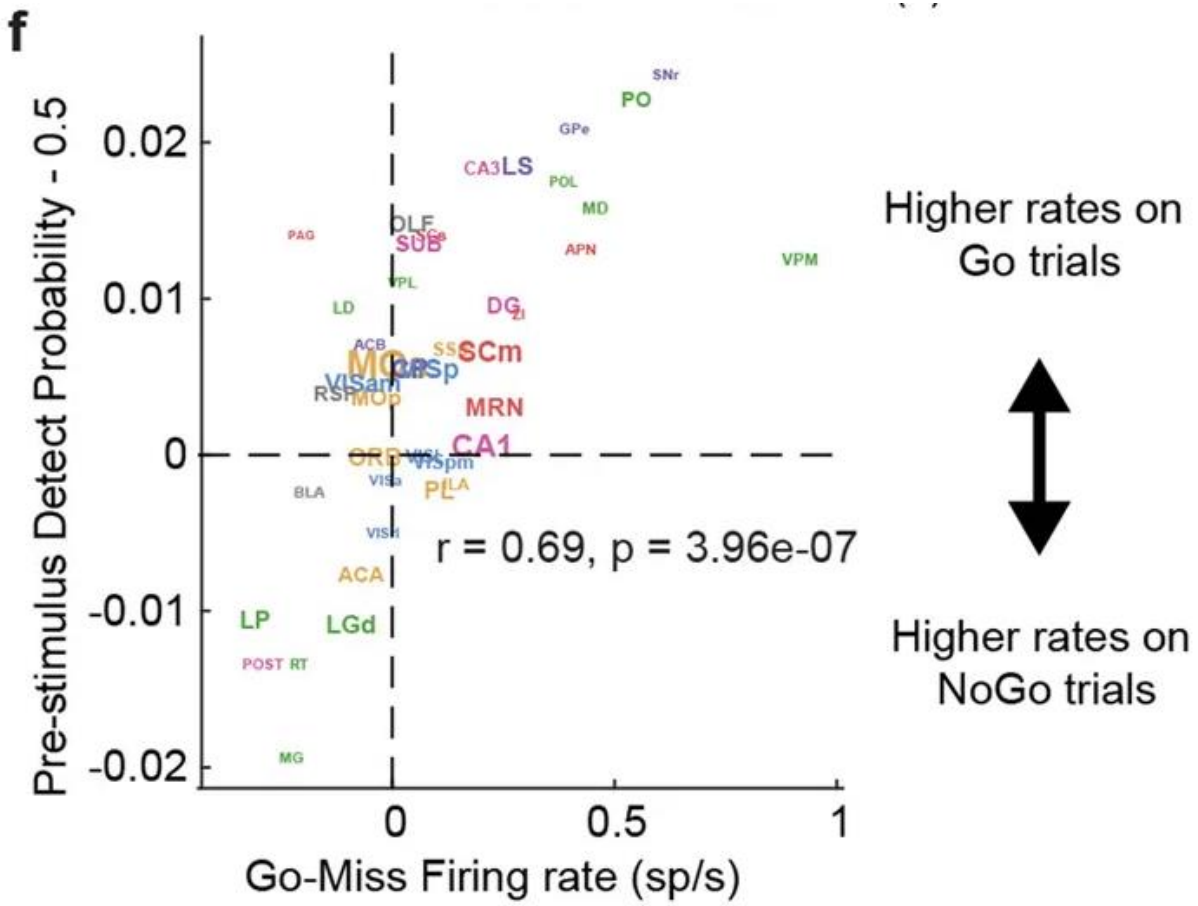
We tend to focus on the area of the shape rather than the height/width separately

Red
+ Green

Major interference

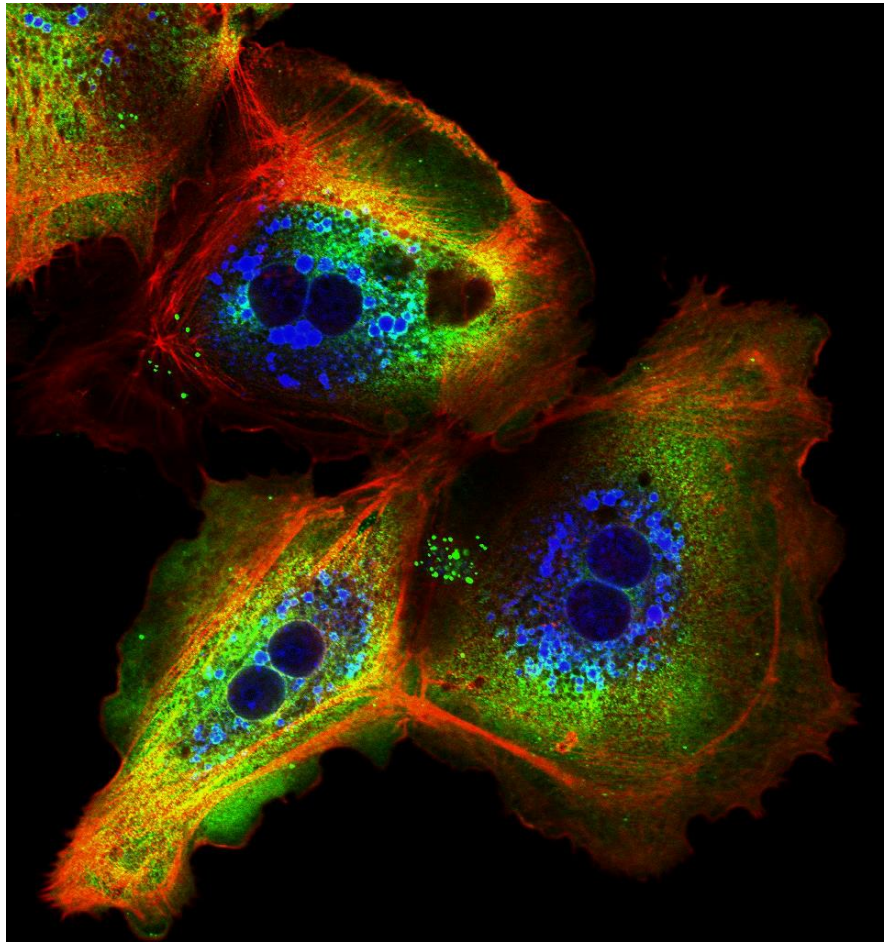
Humans are very bad at separating combined colours

Separability



Higher rates on Go trials

Higher rates on NoGo trials



Other visual cues

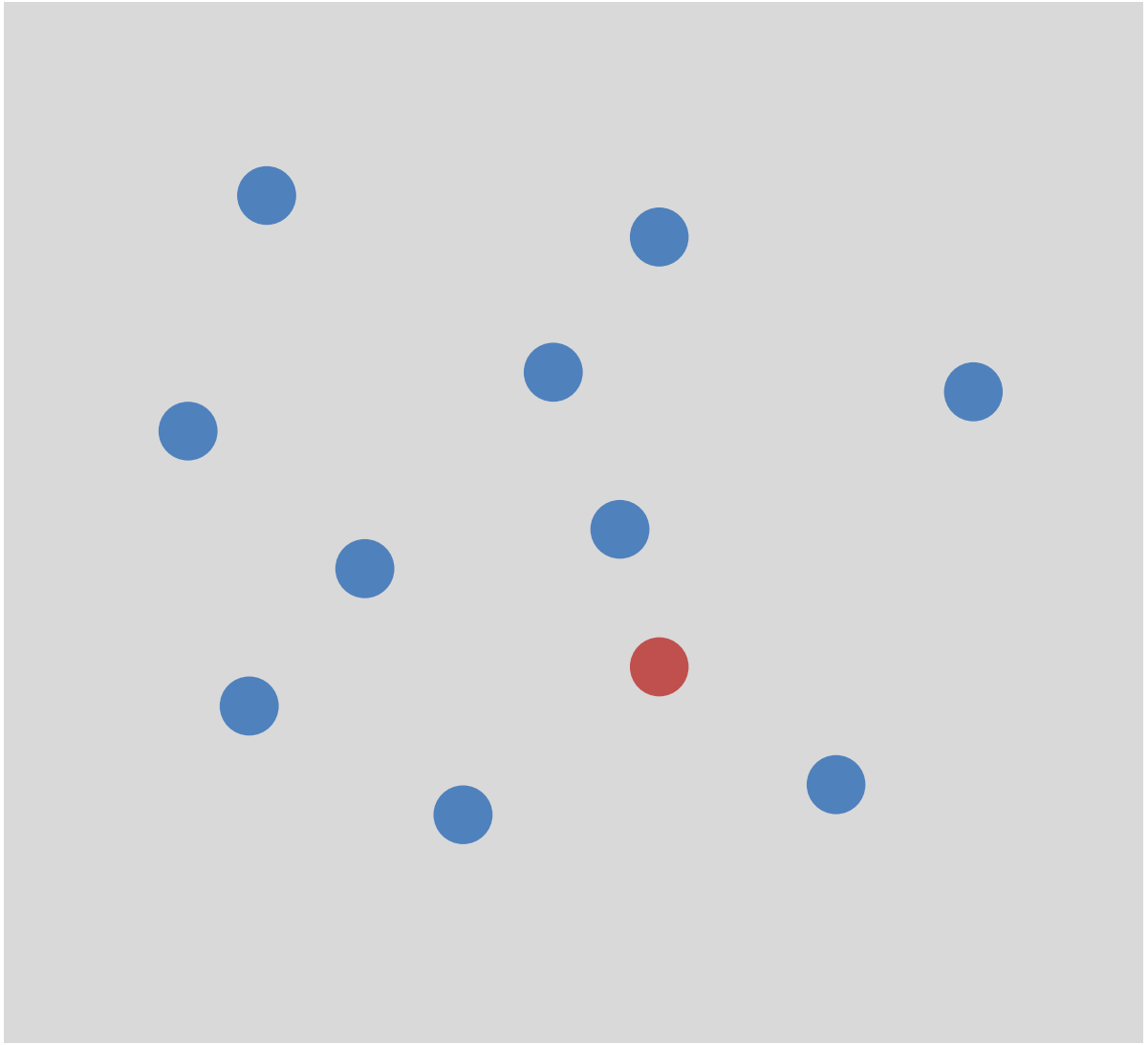
How can you modify your plot to improve its ease of interpretation, without changing the basic data representation?

Other visual cues
Popout

- Sometimes you want to draw people's attention to parts of the plot
- We can use colours or shapes to trigger a 'popout' reaction
- An implicit rather than explicit cue

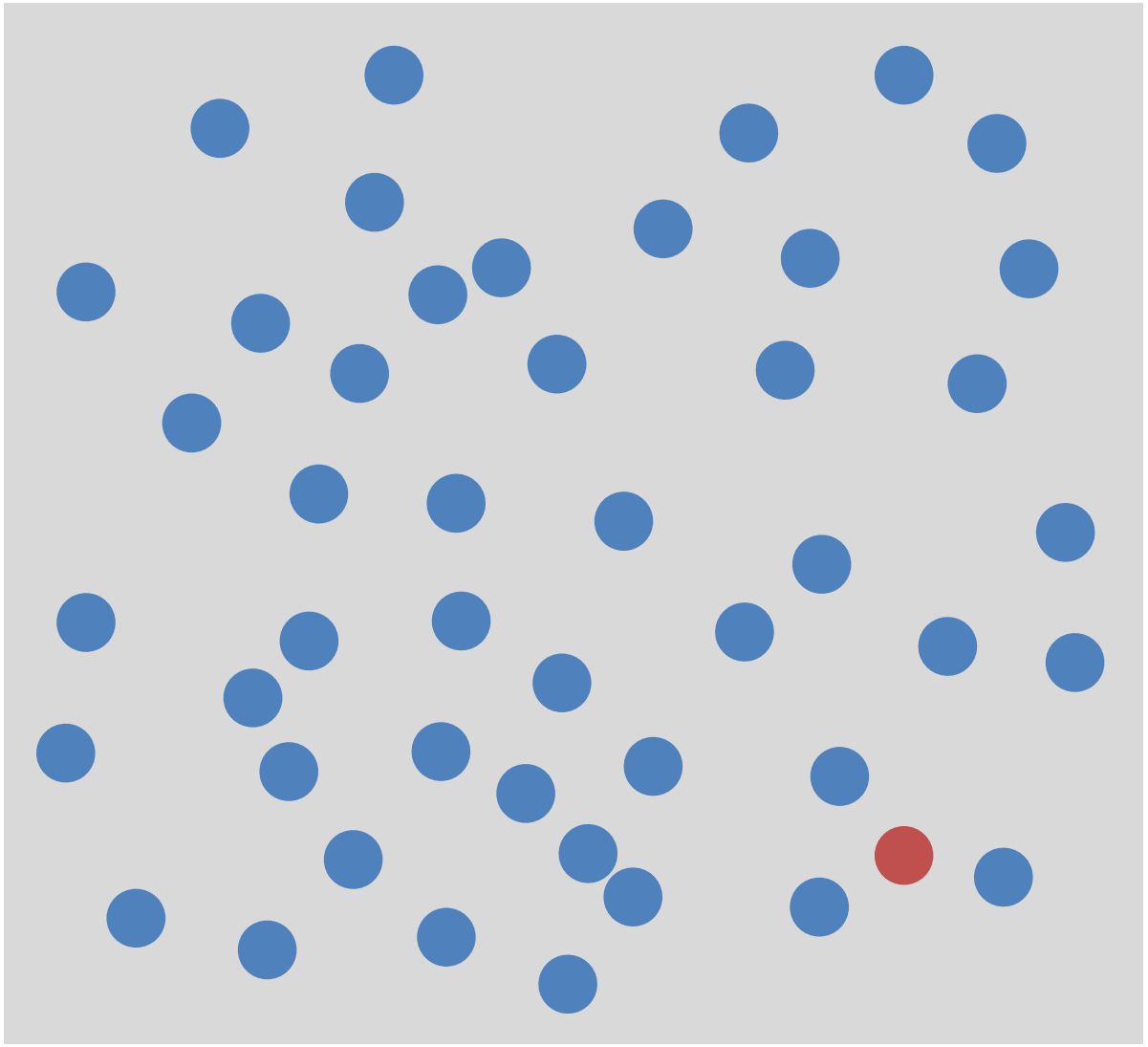
Popout

(find the red circle)



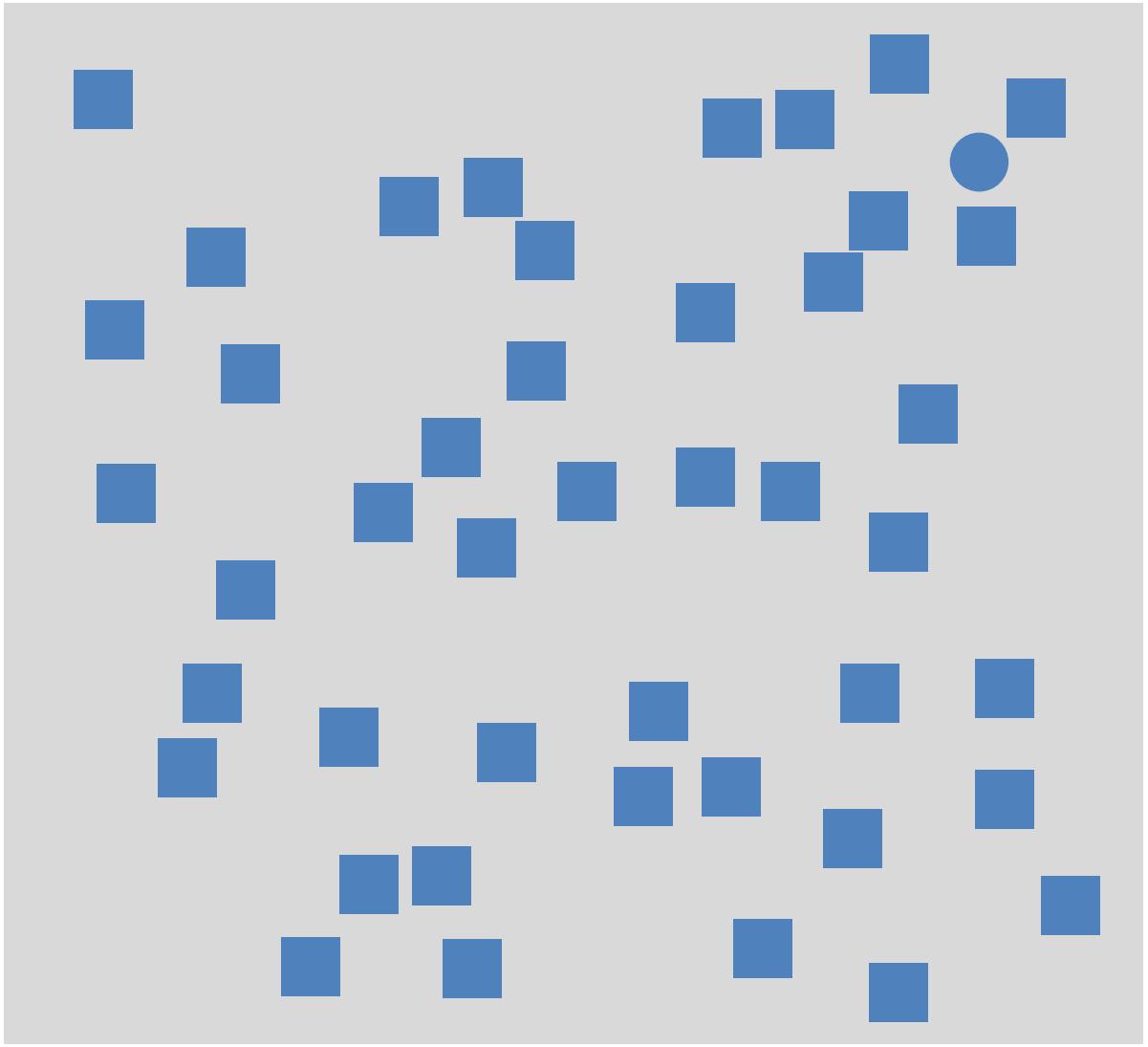
Popout

Speed of identification is independent of the number of distracting points



Popout

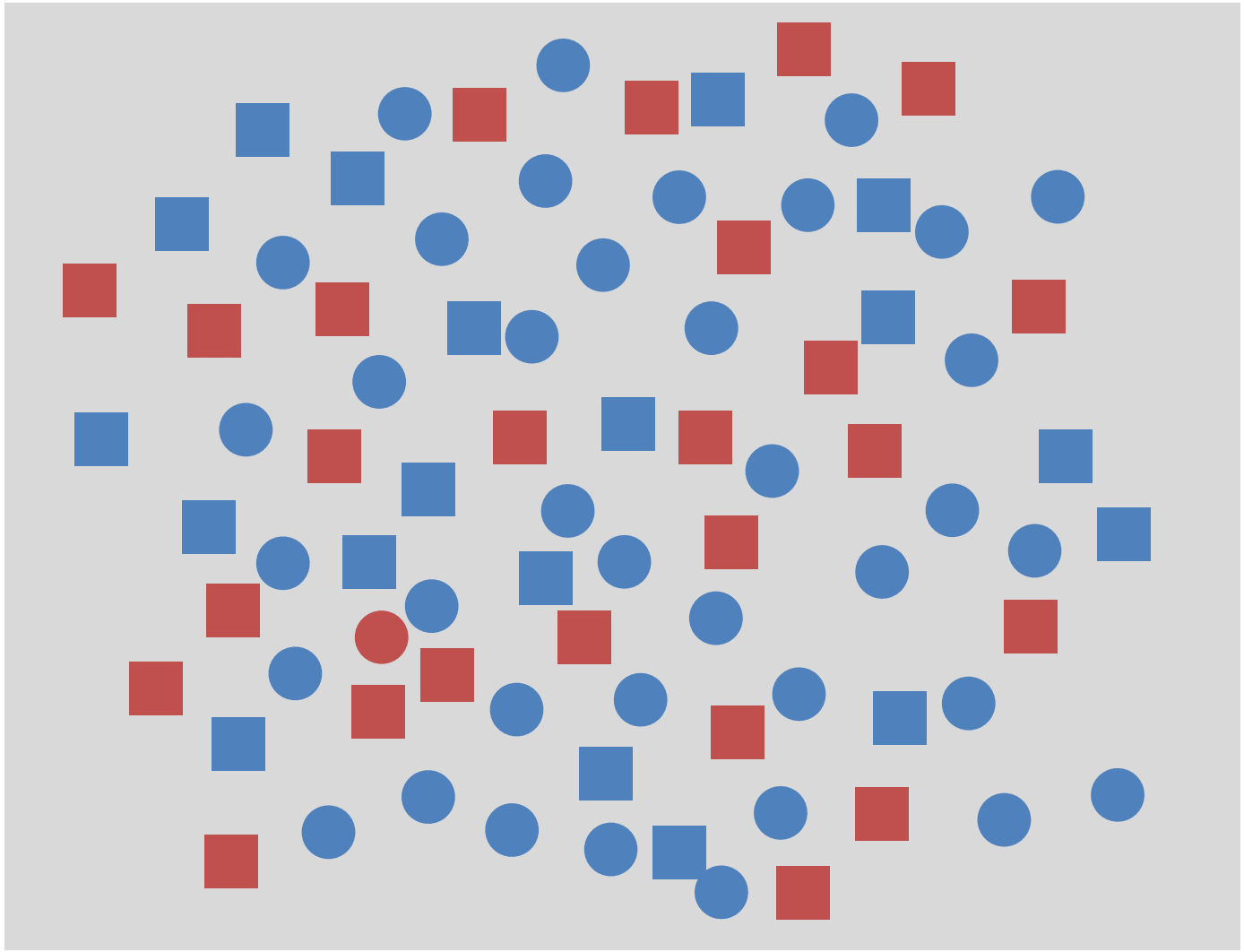
Colour pops out more than shape



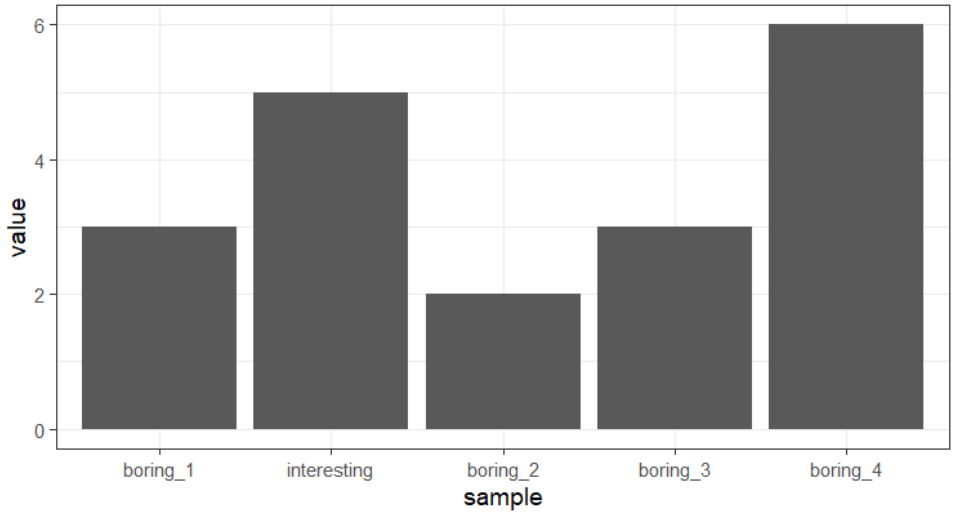
Popout

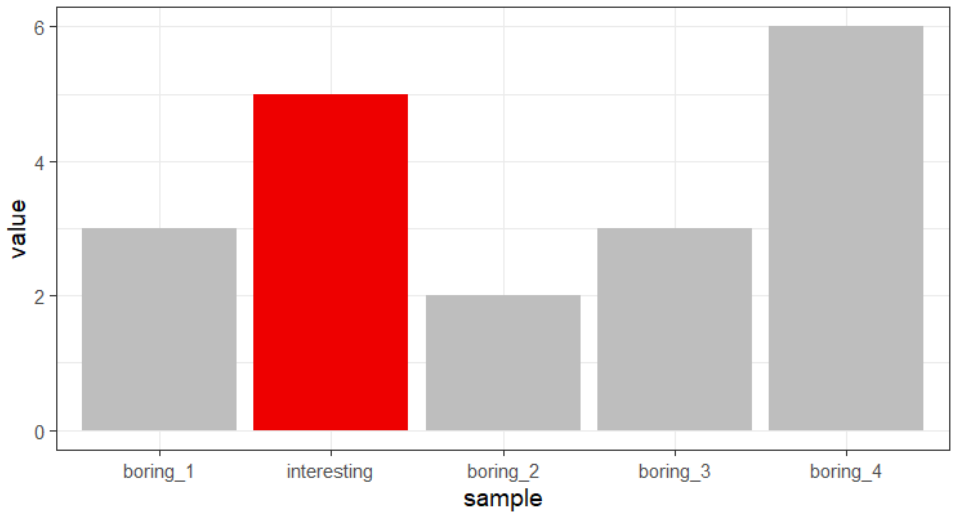
Mixing channels removes the effect

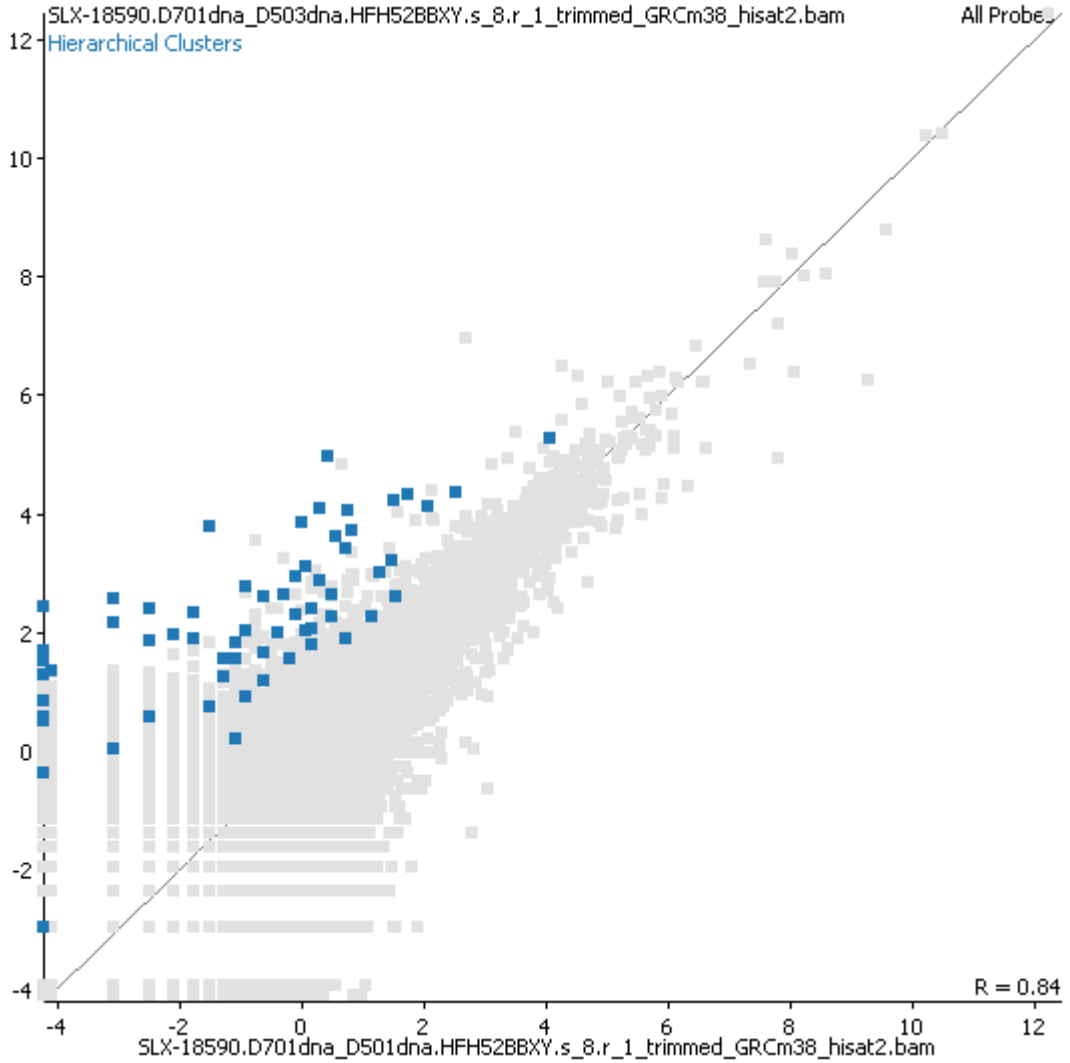
(Find the red circle)



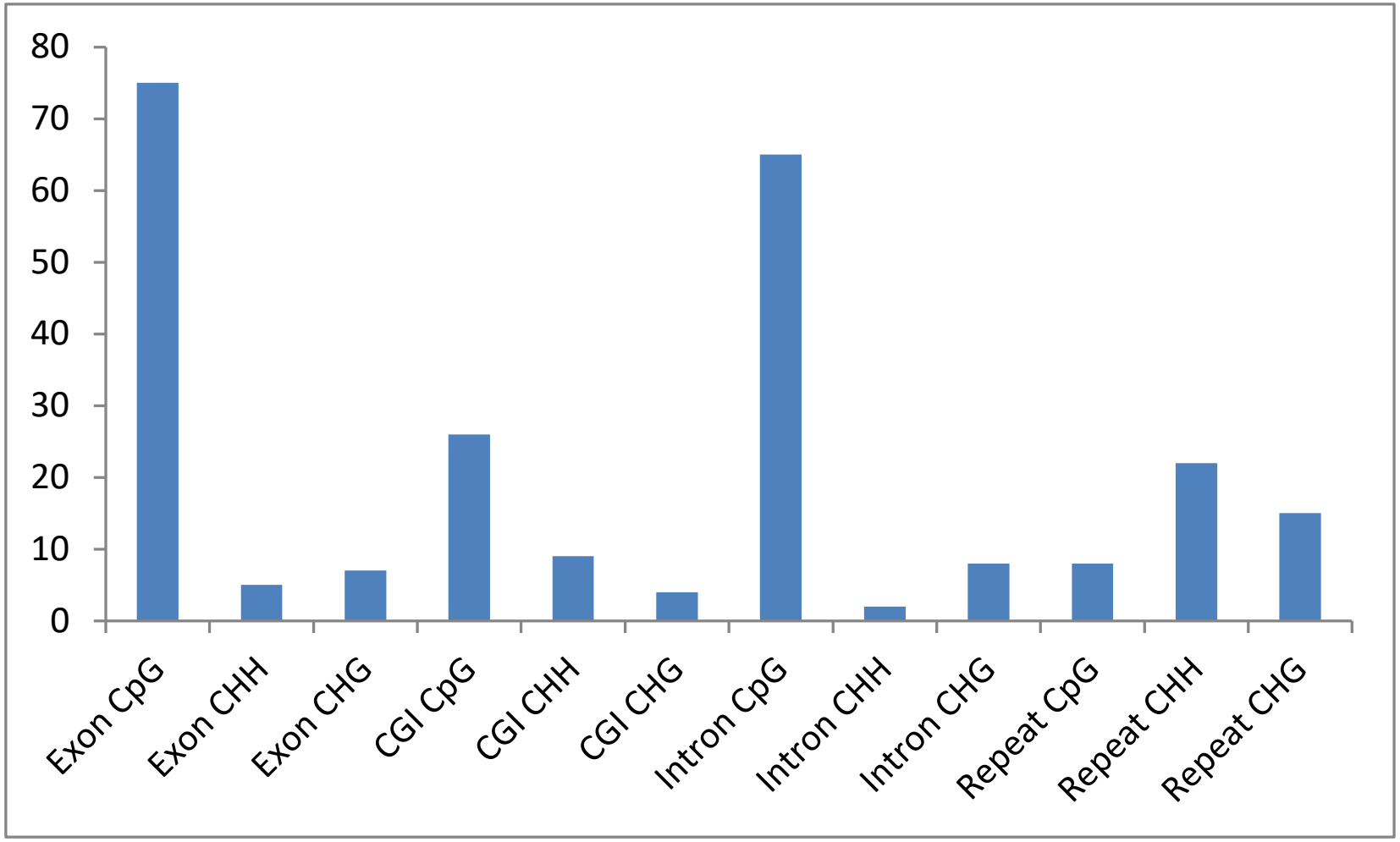
Popout Examples



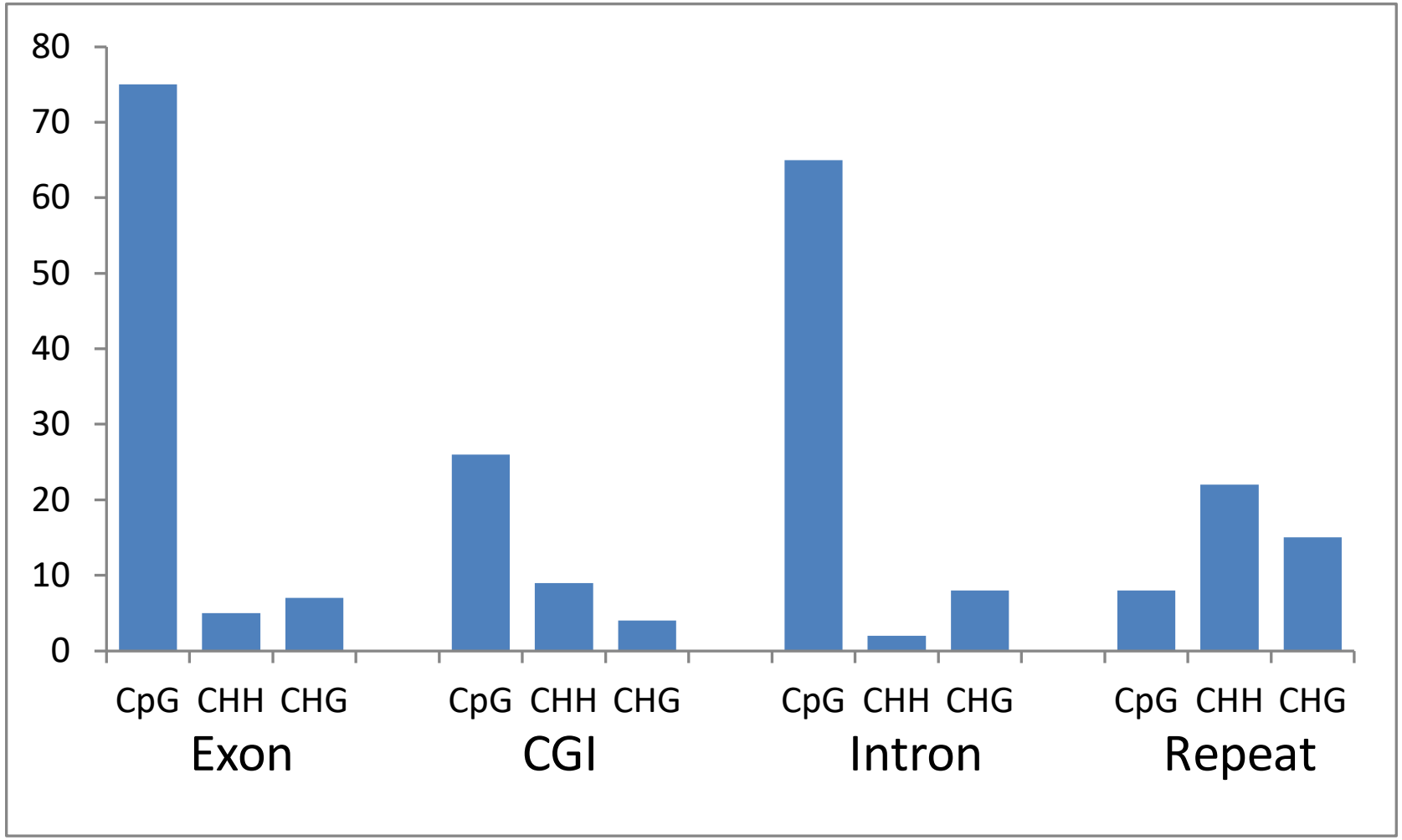




Other visual clues

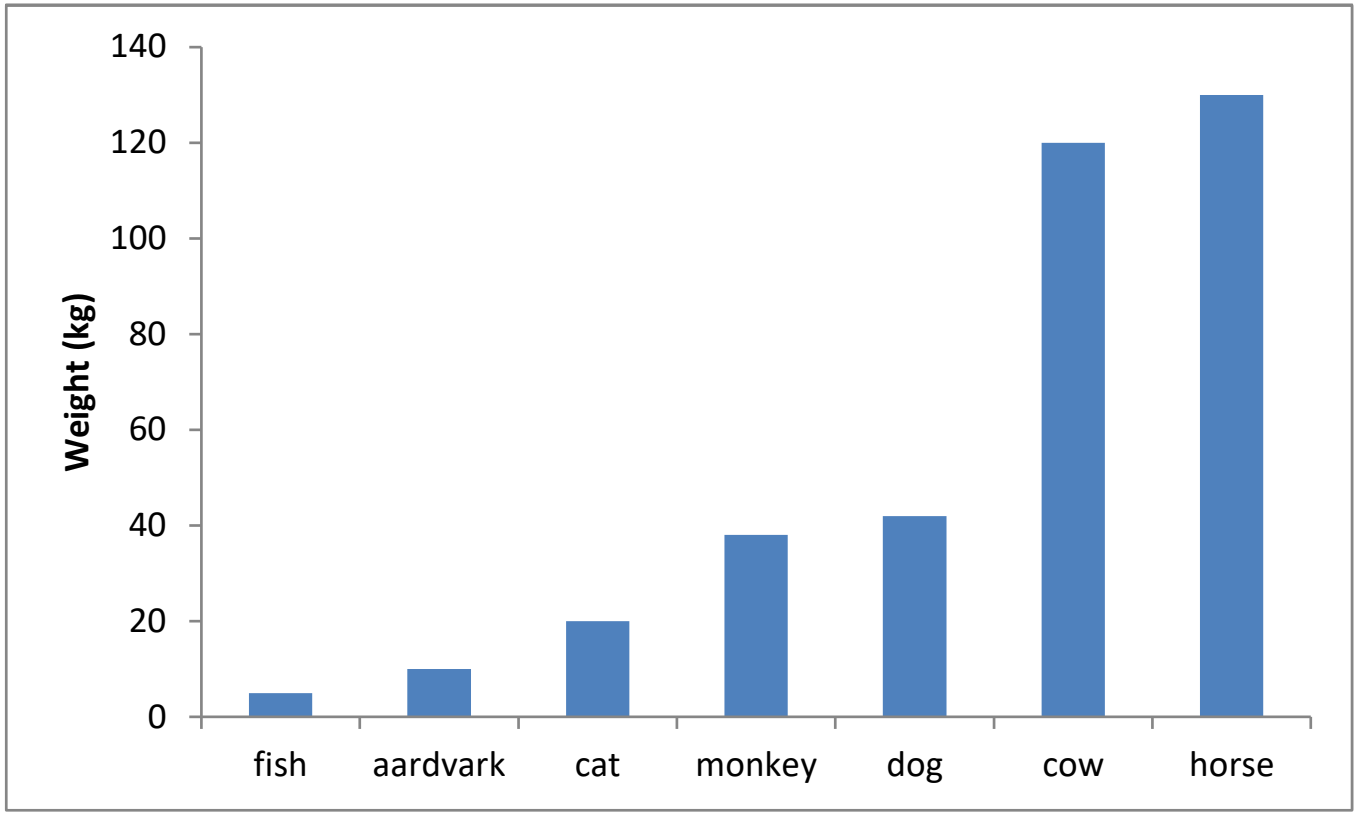


Grouping


80
70
60
50
40
30
20
10
0
CpG CHH CHG
Exon
CpG CHH CHG
CGI
CpG CHH CHG
Intron
CpG CHH CHG
Repeat

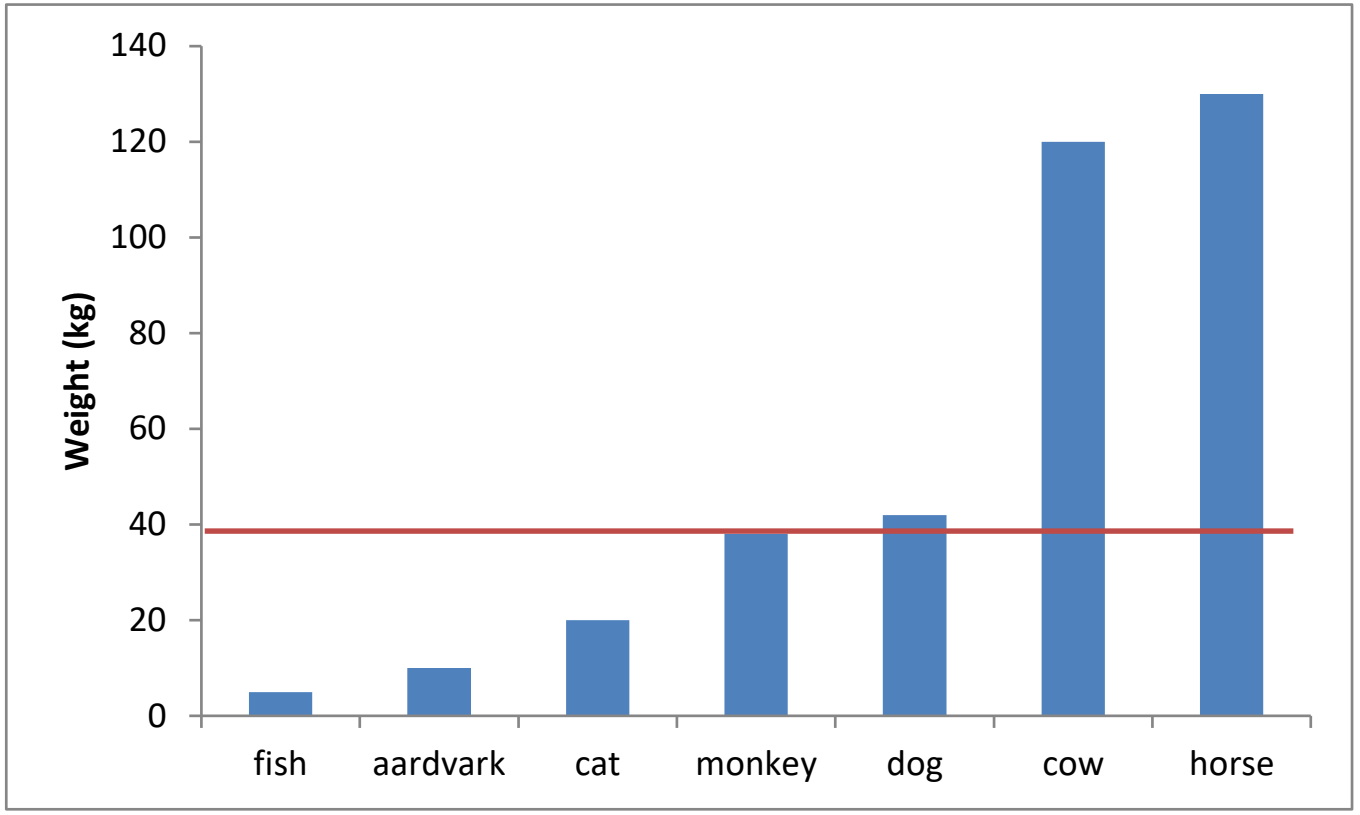
Other visual clues

- Is a monkey heavier than a dog?



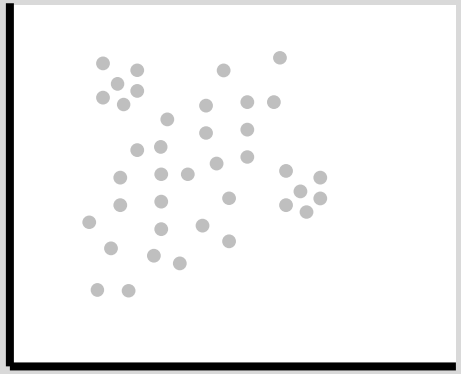
Other visual clues

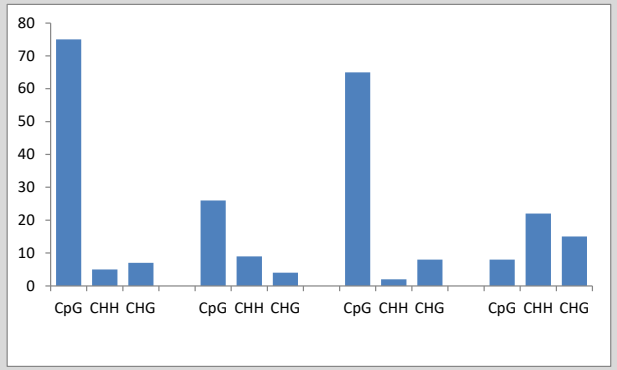
- Is a monkey heavier than animal X?



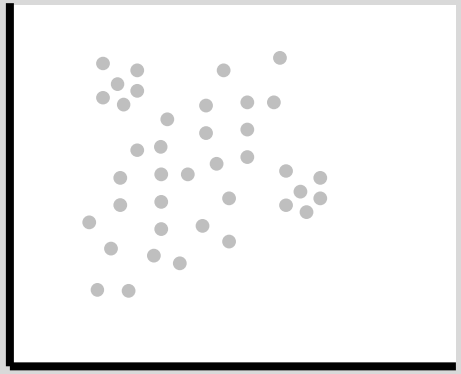
Containment / Linking

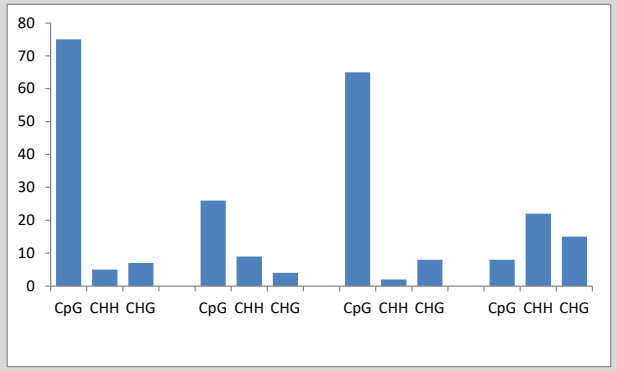
Wild Type



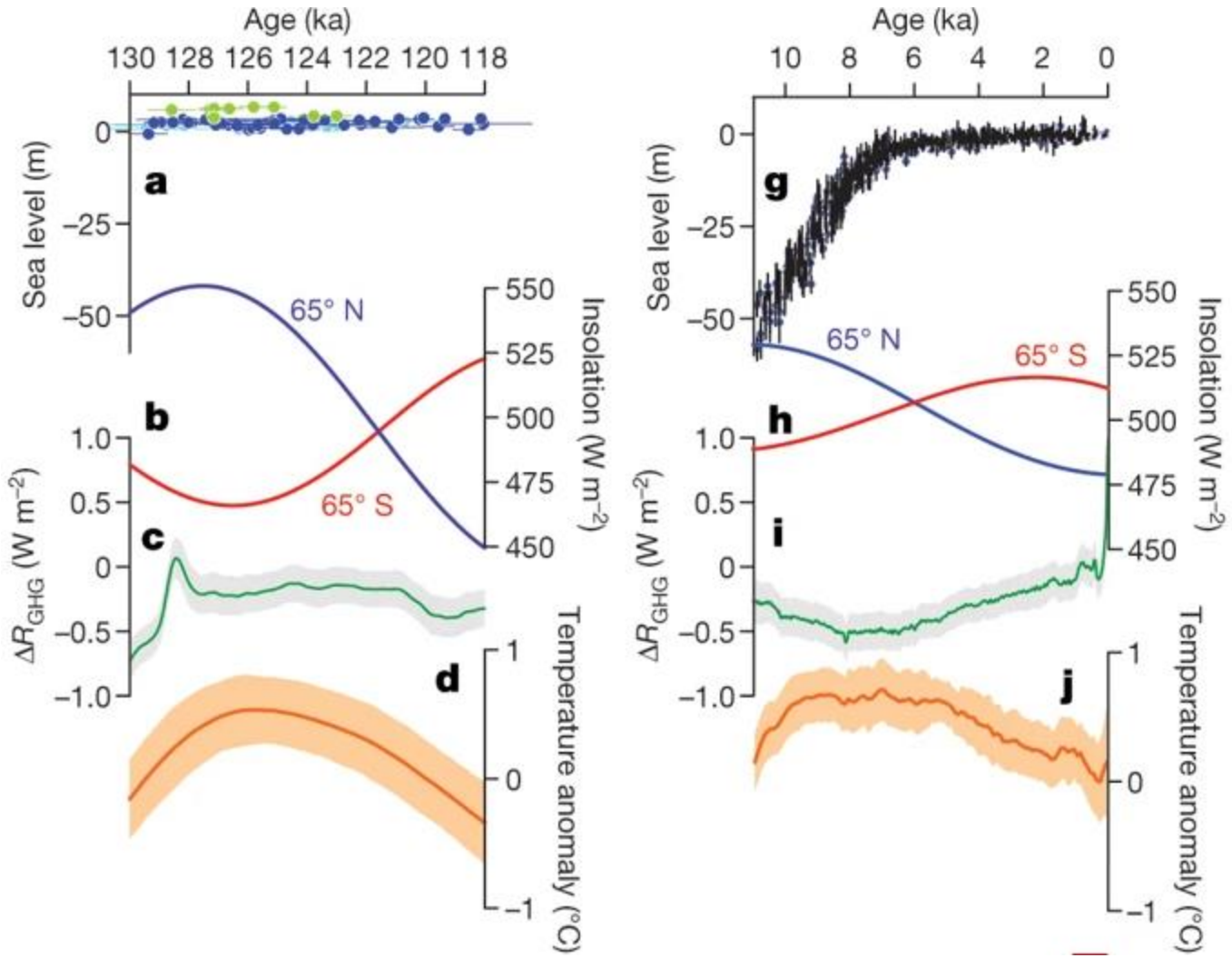


Mutant





Containment / Linking



How do you know if your figure is working?

Validation

- Always try to validate plots you create
- You have seen your data too often to get an unbiased view
- Show the plot to someone not familiar with the data
  - What does this plot tell you?
  - Is this the message you wanted to convey?
  - If they pick multiple points, do they choose the most important one first?

Exercise

You will be given a series of (not very good) plots to validate. Try to think what message the plot is trying to convey and whether it is doing so effectively.

Work out how you would choose to represent the data if you don’t like the way it’s presented now.

Making effective use of common plot types

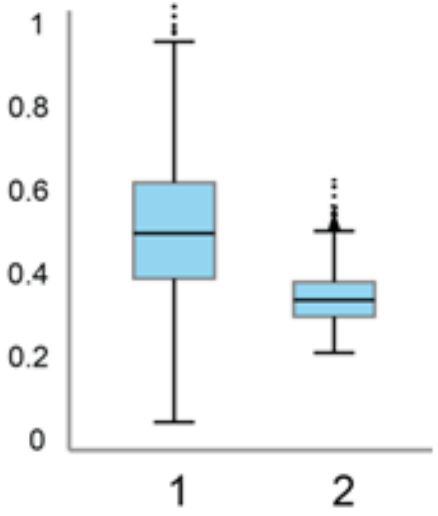
Anne Segonds-Pichon
Simon Andrews
Phil Ewels

simon.andrews@babraham.ac.uk

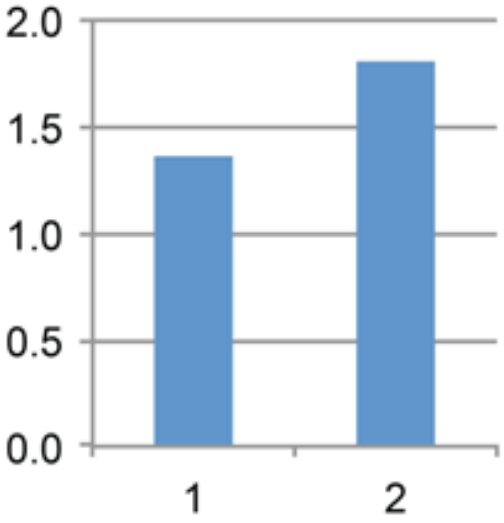
Babraham Bioinformatics

Types of plot

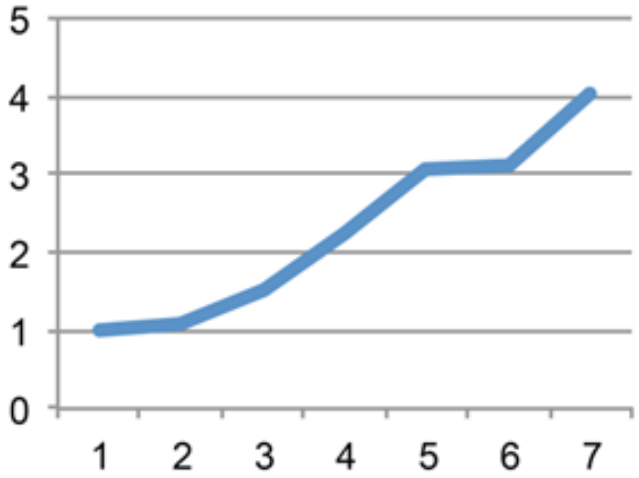
Things you can illustrate



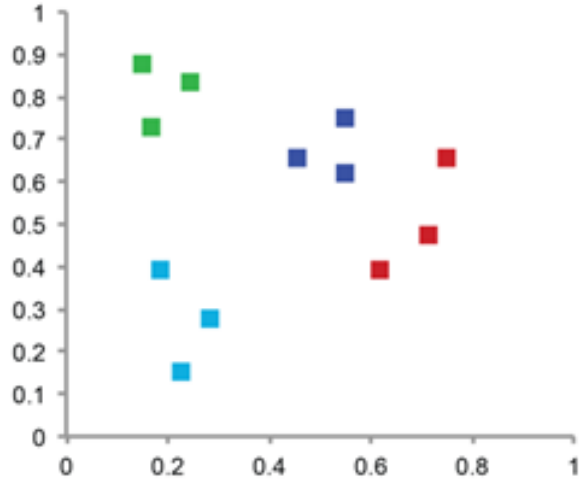
Distribution



Comparison



Relationship

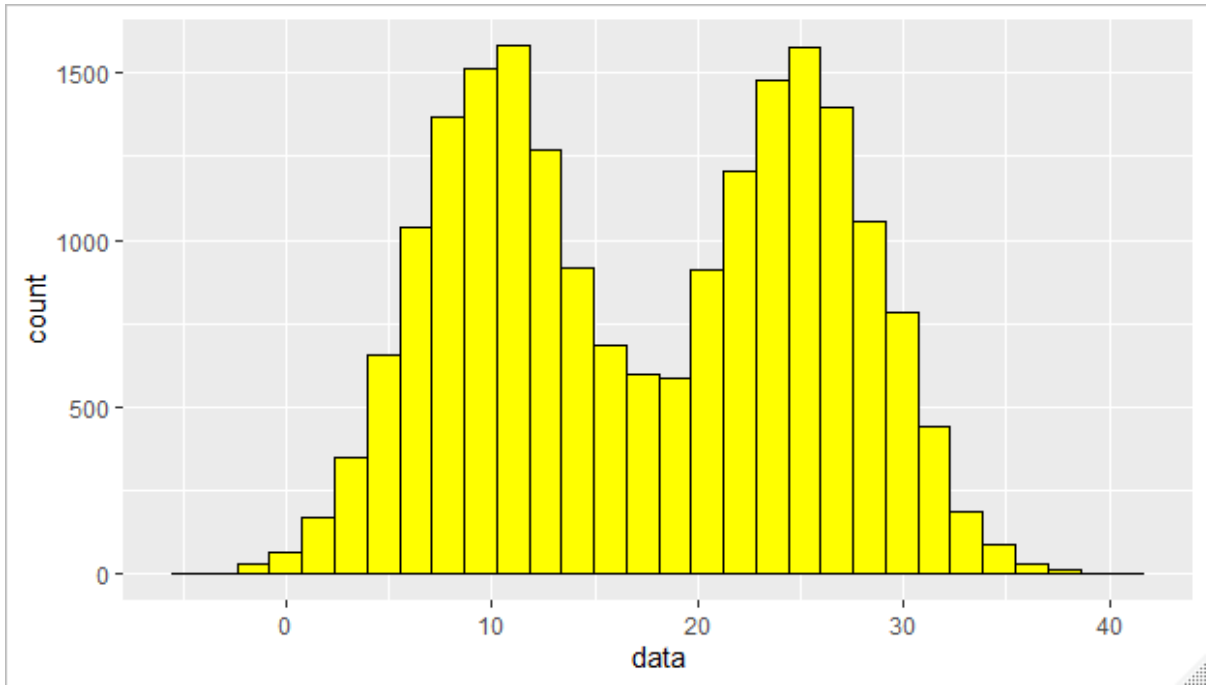


Composition

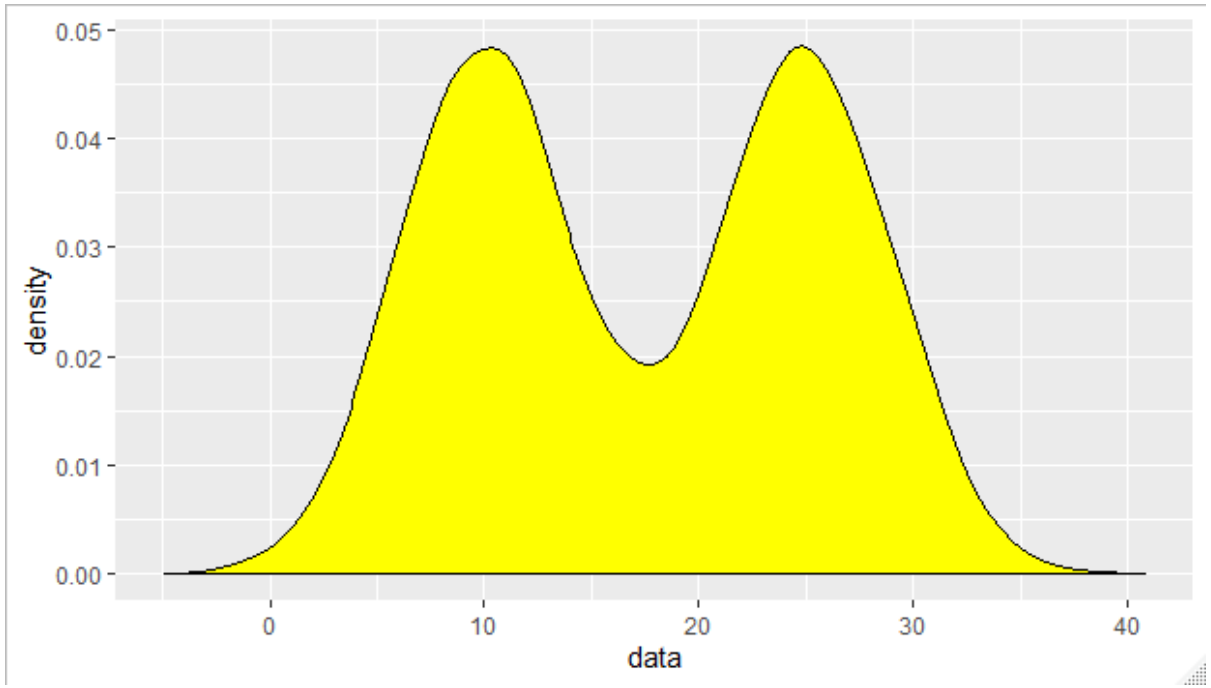
Distributions

Representing Distributions
Single Samples

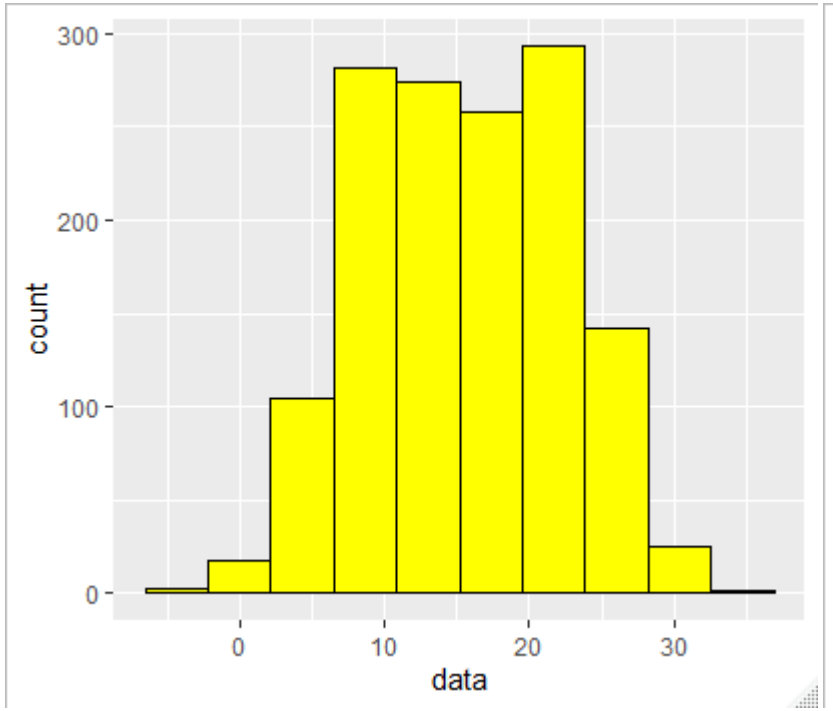
Histograms

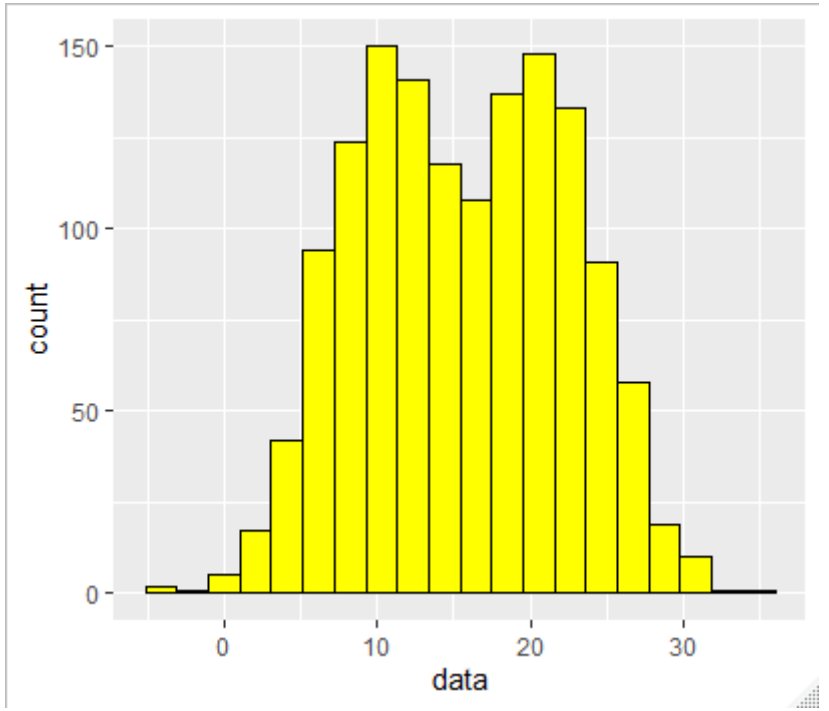


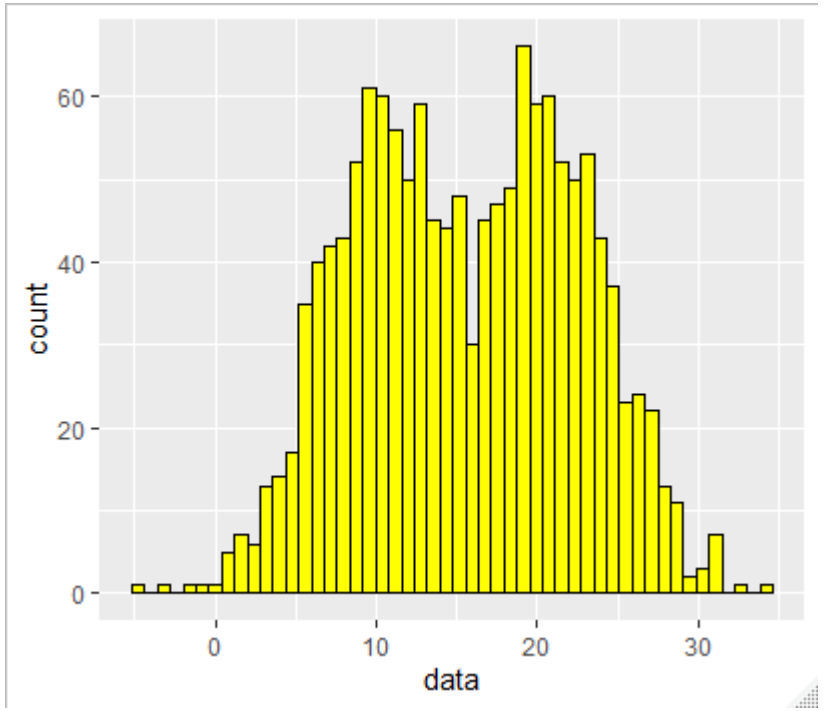
Density Plots



Representing Distributions
Single Samples - Bandwidth

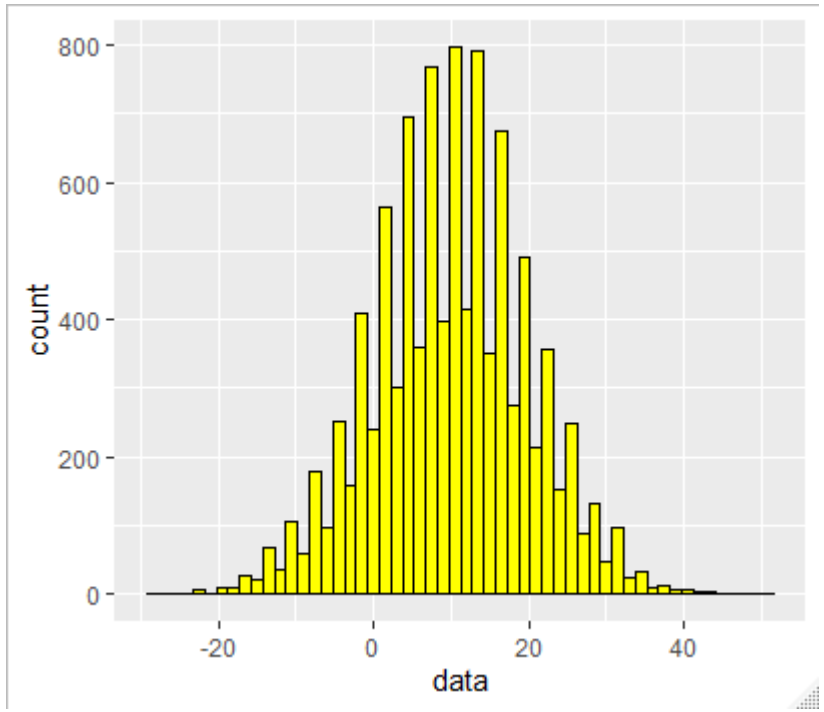




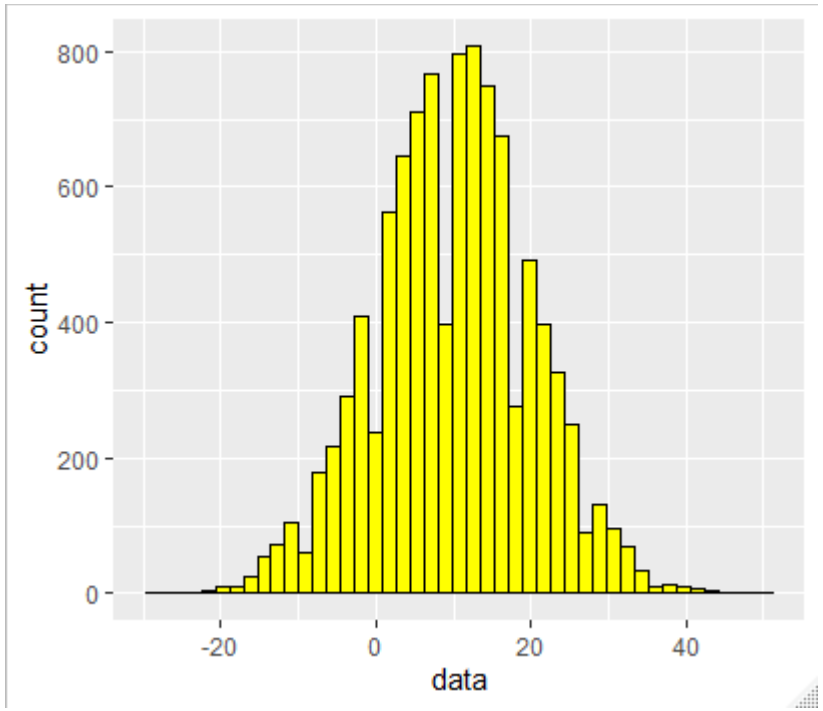


Representing Distributions
Single Samples – Discontinuous data

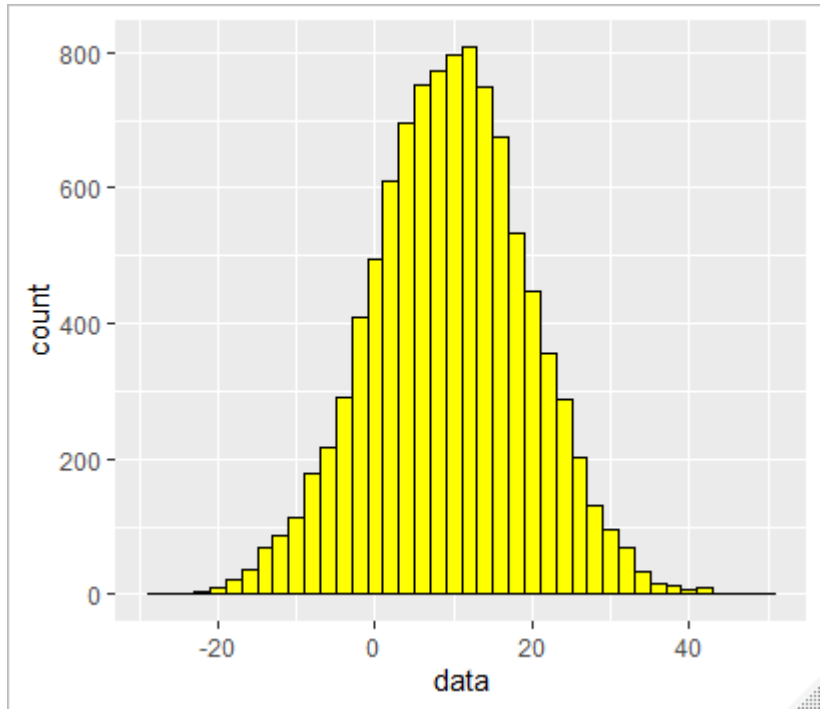
Plotting Integer Data



1.5

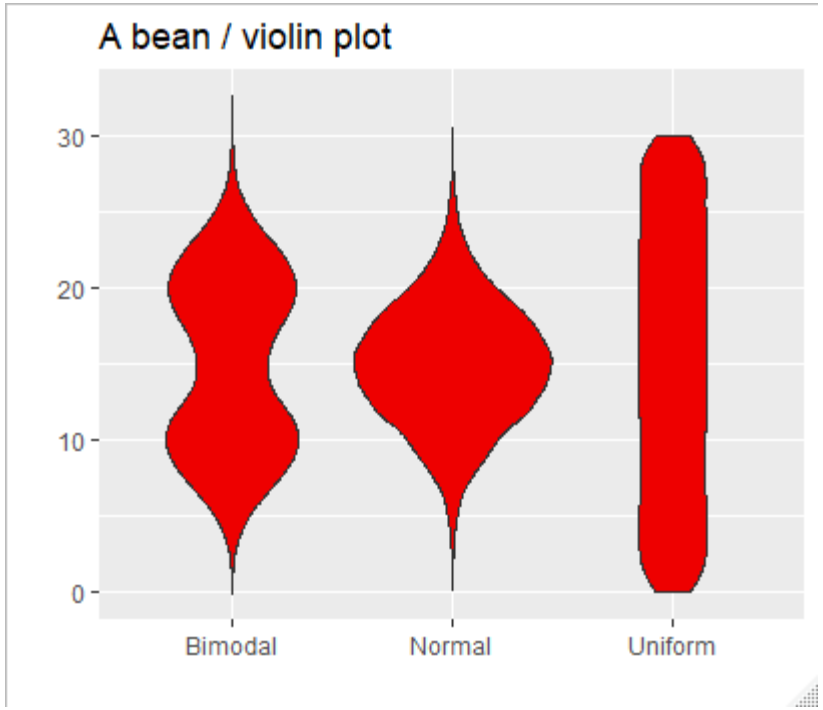


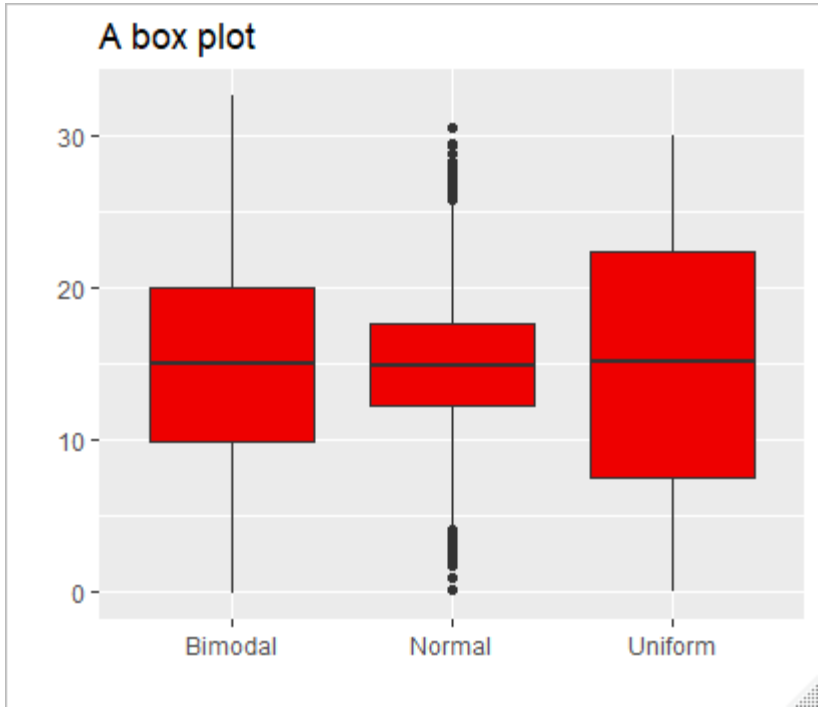
1.8

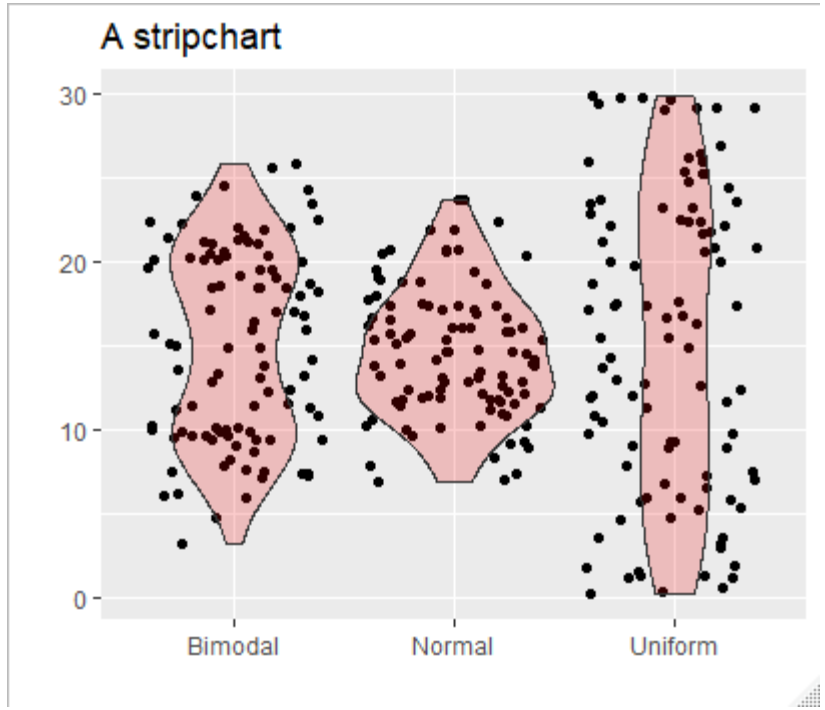


2

Representing Distributions
Multiple Samples

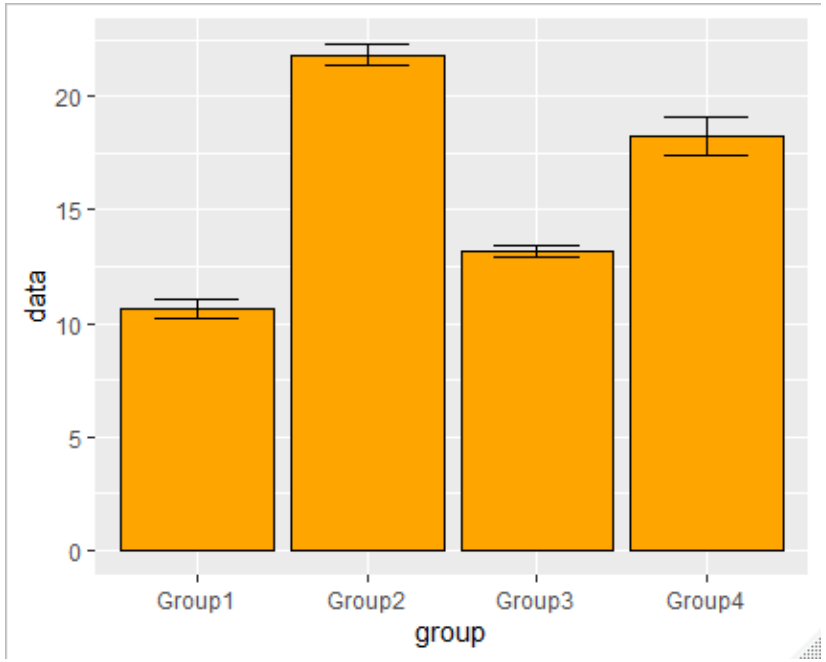


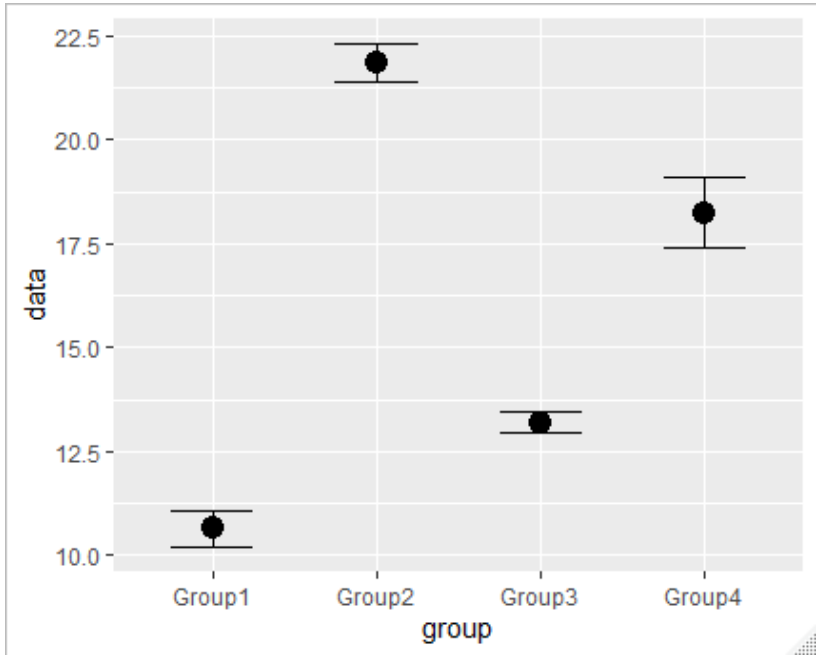


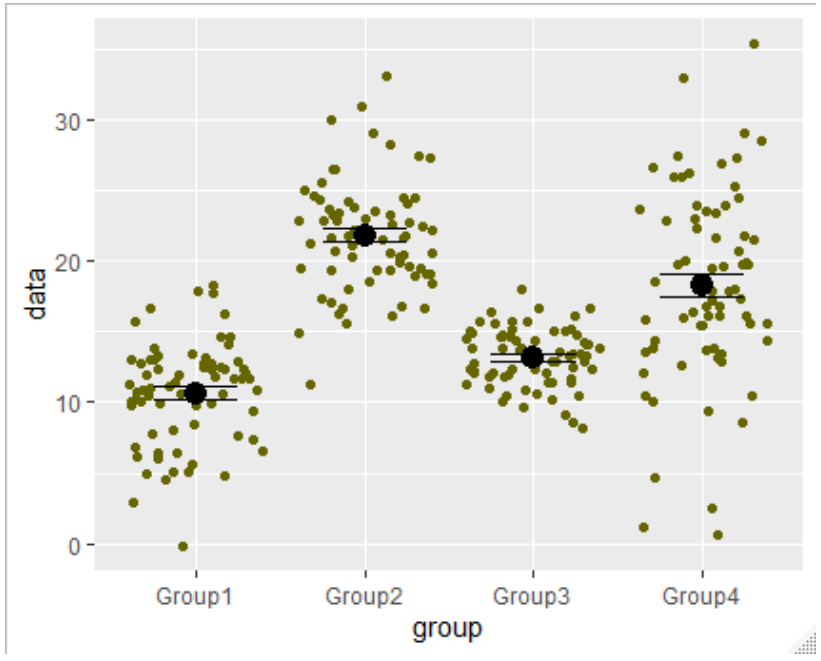


Comparisons

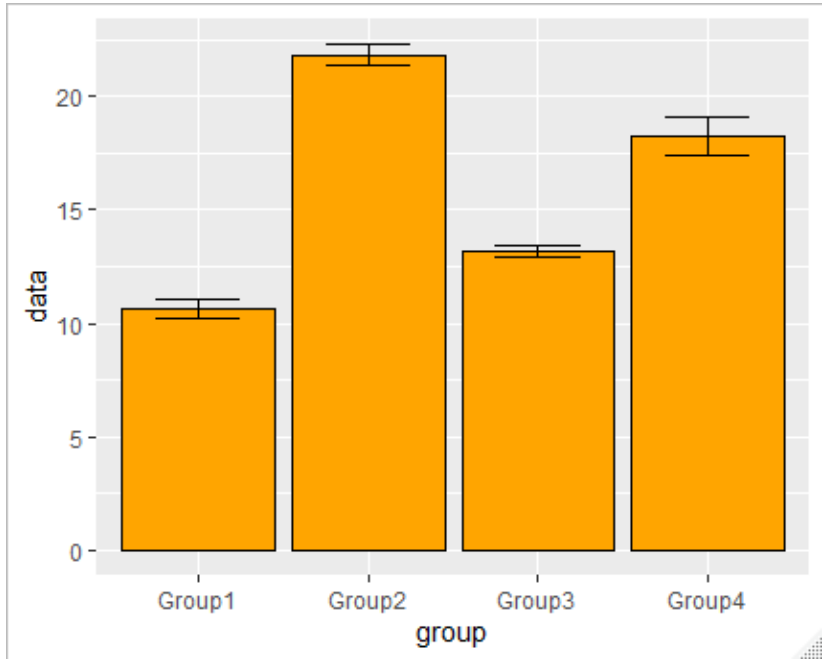
Comparisons



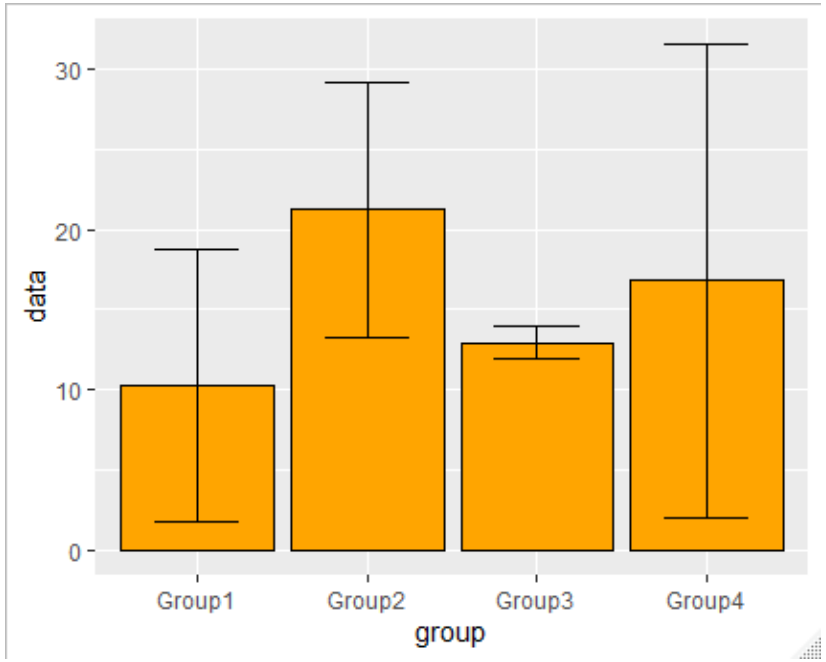




Error Bars

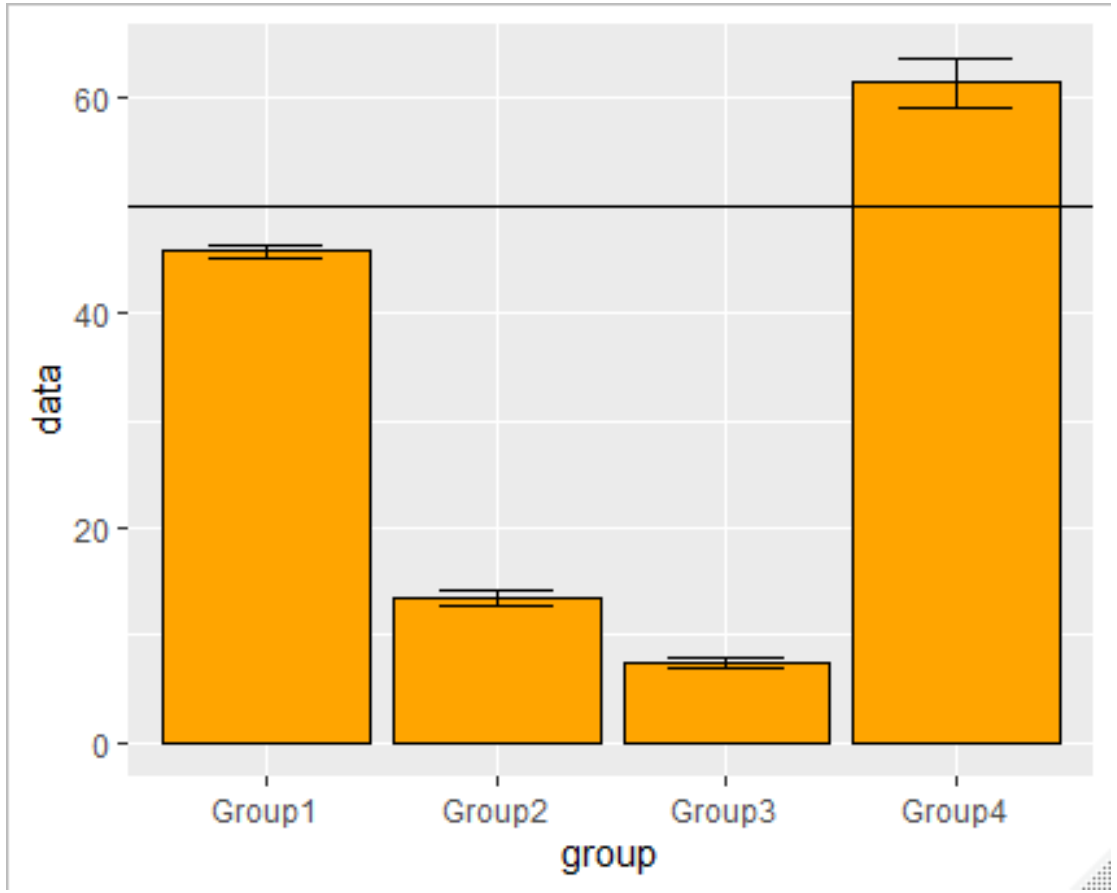


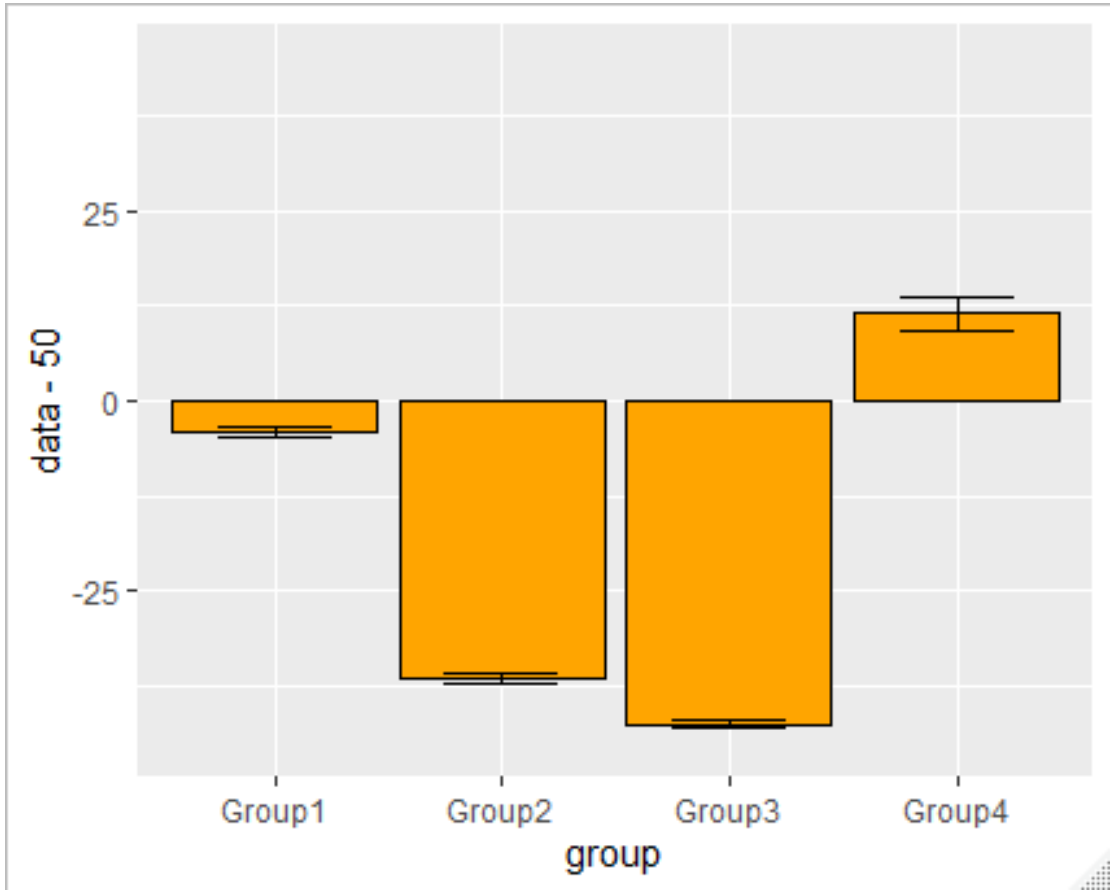
- Standard Error of Mean (SEM)
- How accurately is the mean calculated
- Gets smaller with increased data
- Good when comparing means



- Standard Deviation (SD)
- How well does the mean summarise the data
- No systematic change with increased data
- Good when comparing variability

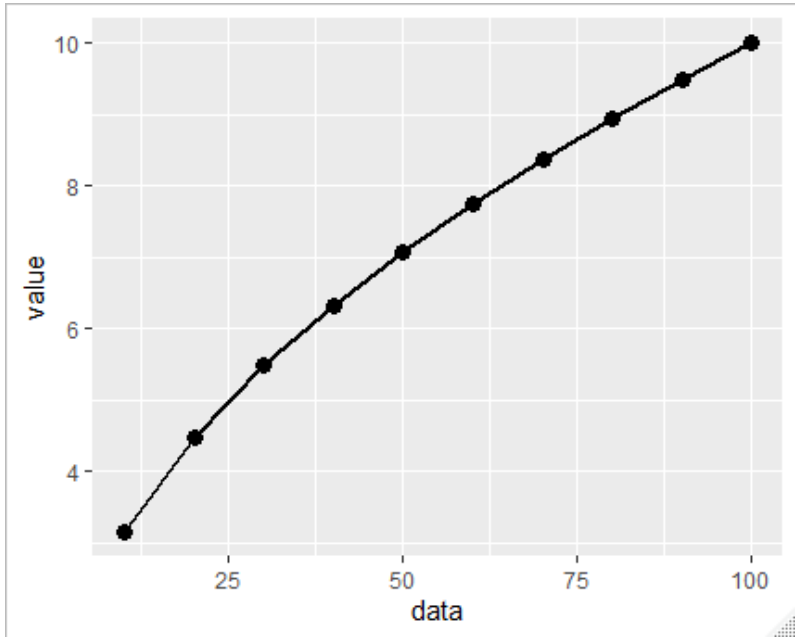
Setting a suitable baseline

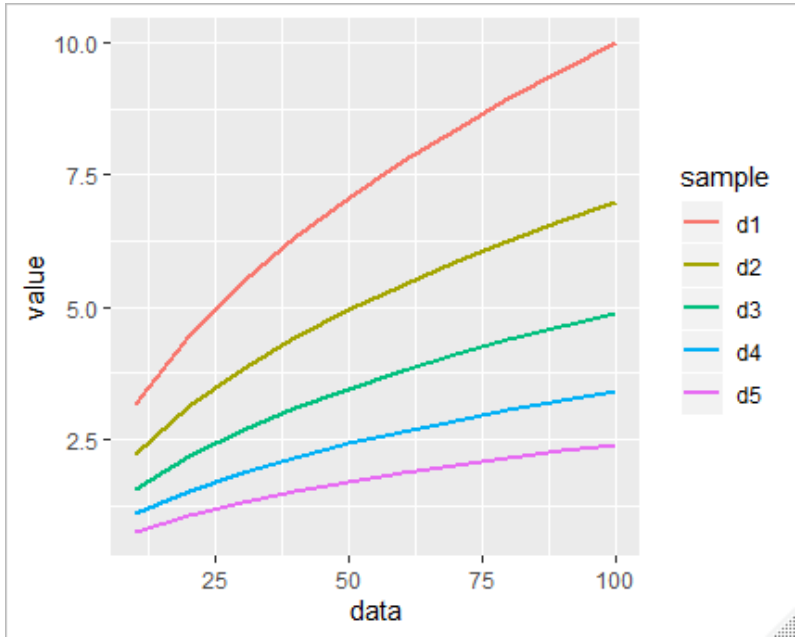


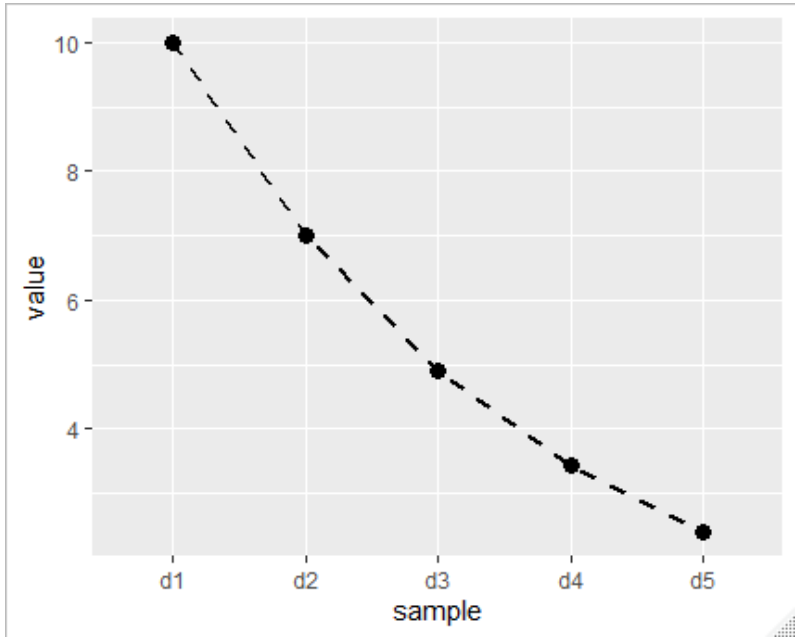


Relationships

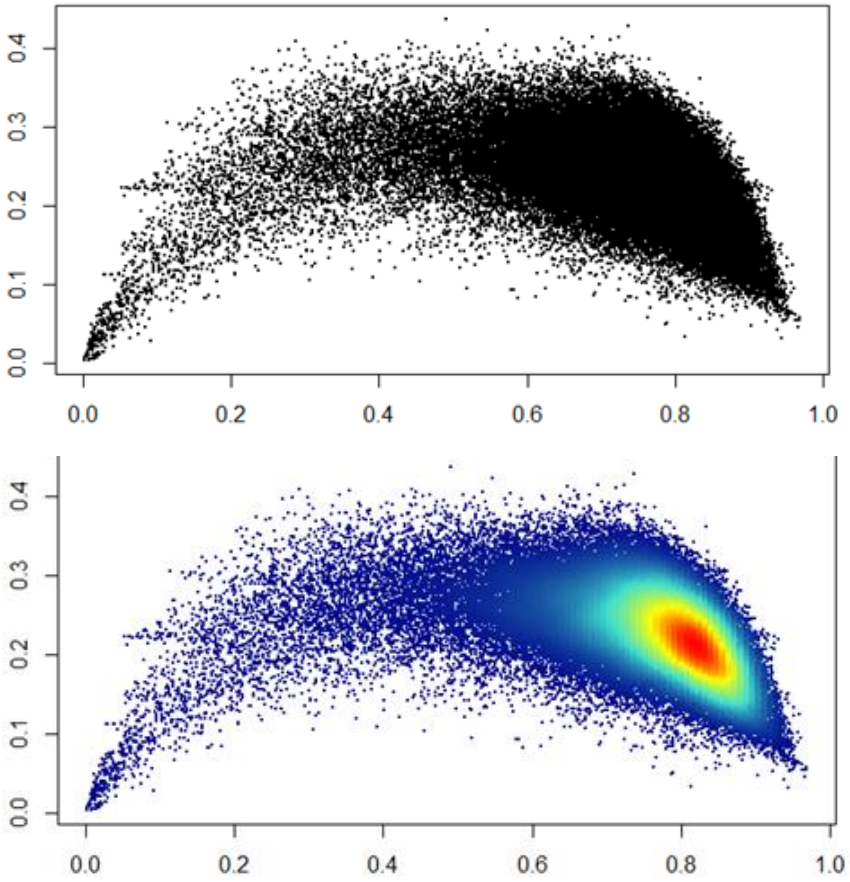
Relationships – Line Graphs

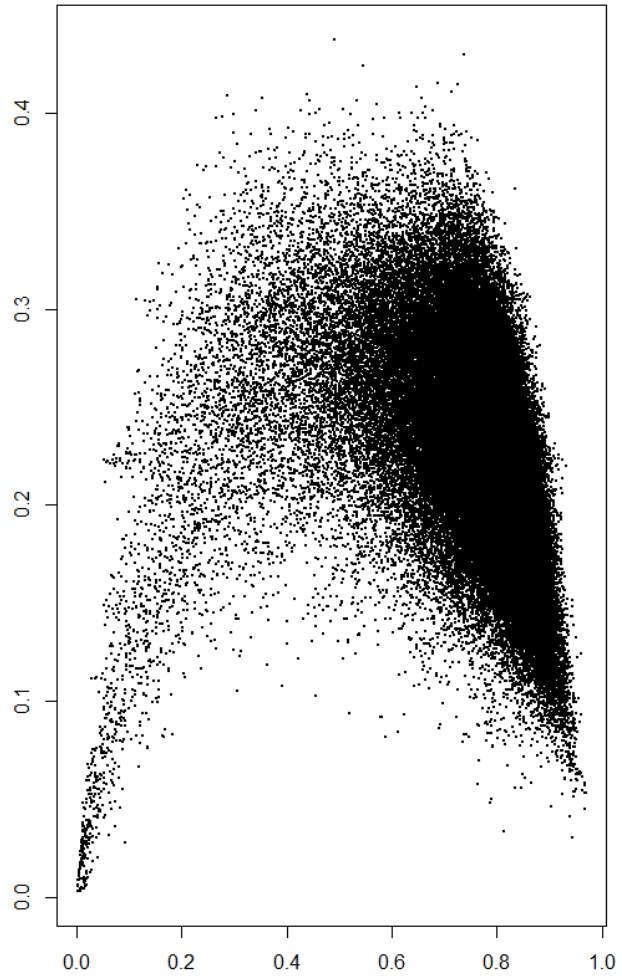






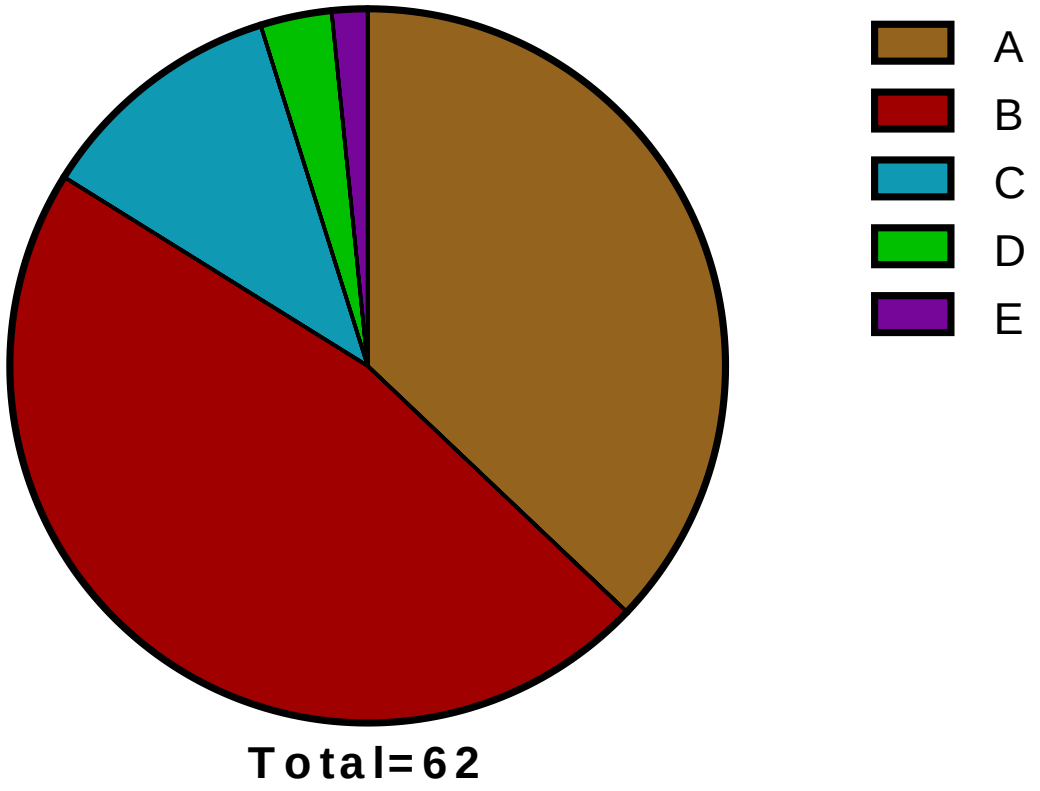
Relationships - Scatterplots

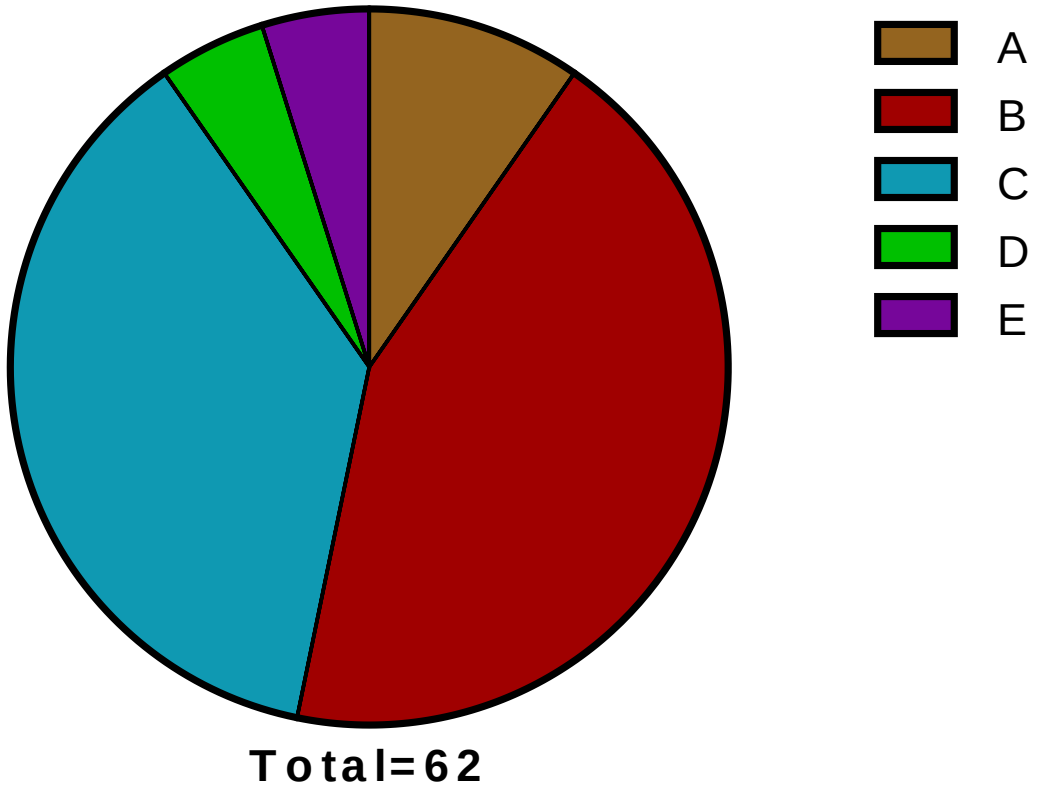




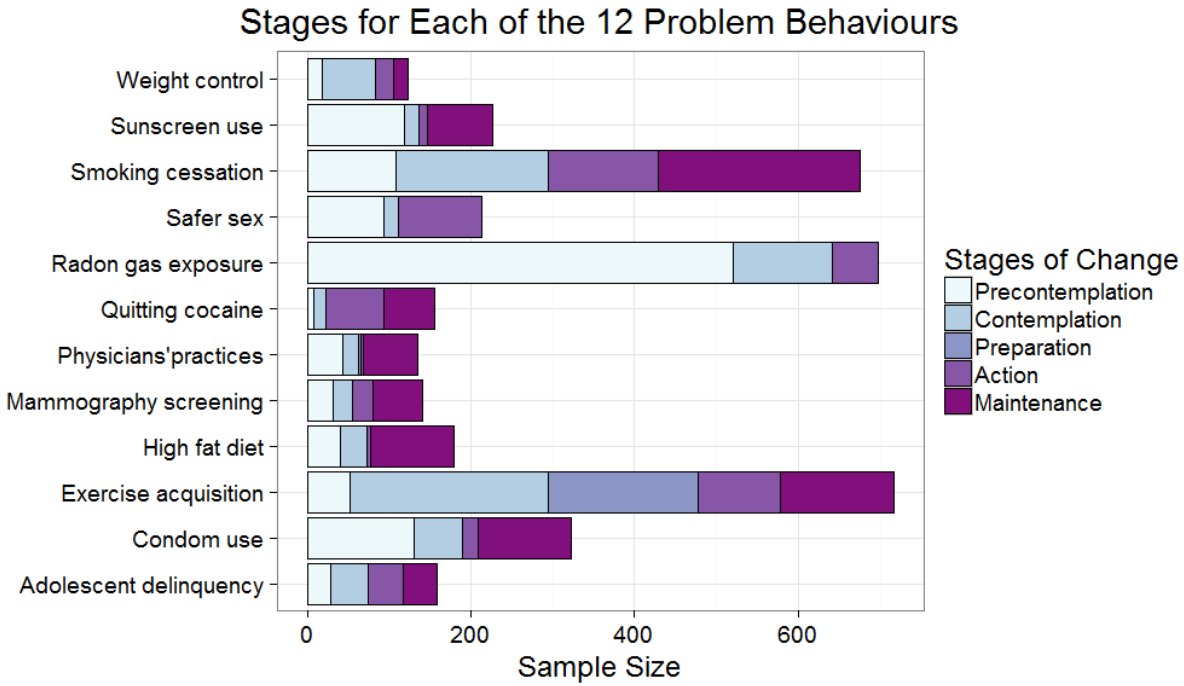
Composition

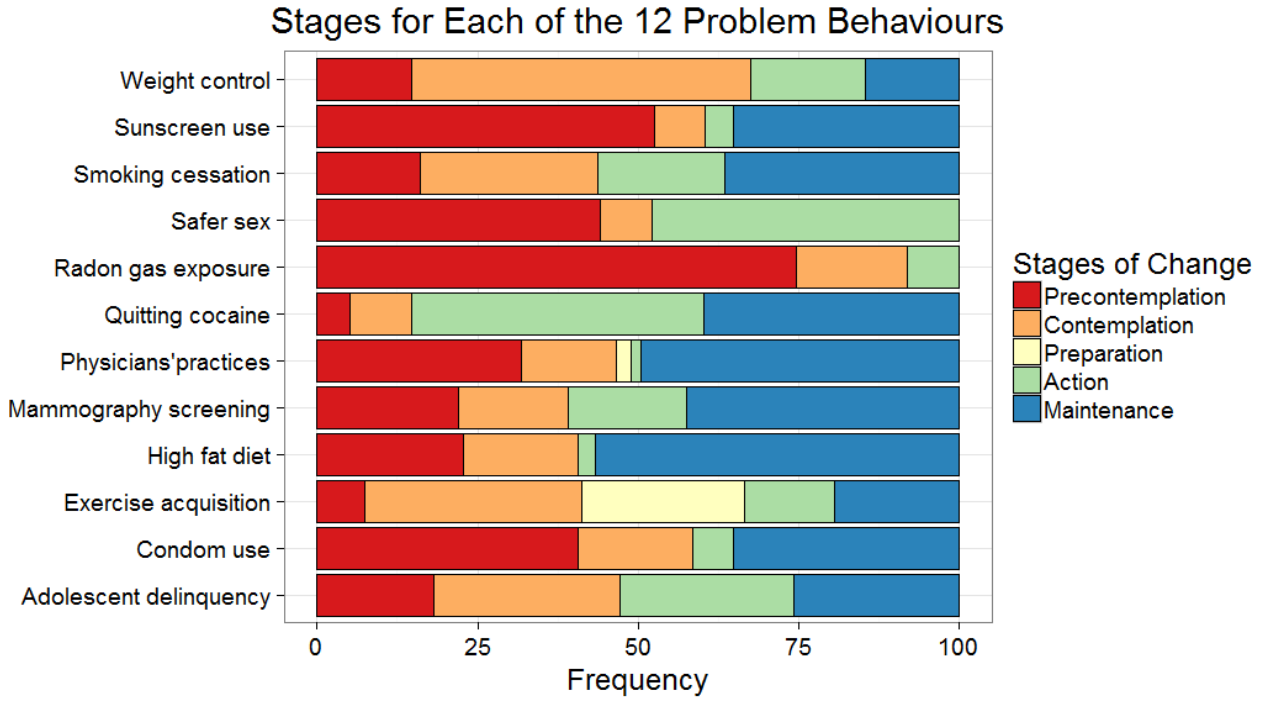
Pie Charts



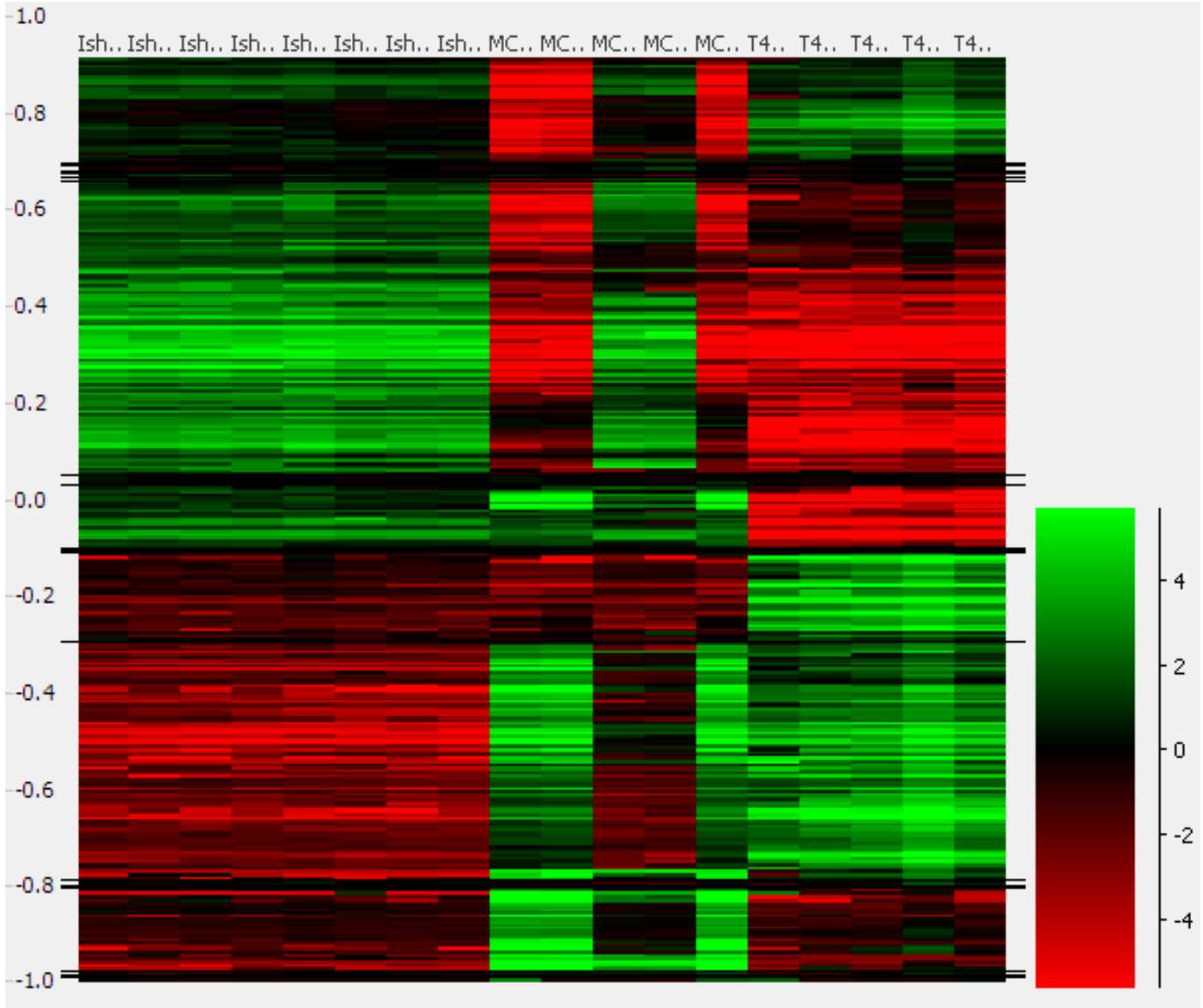


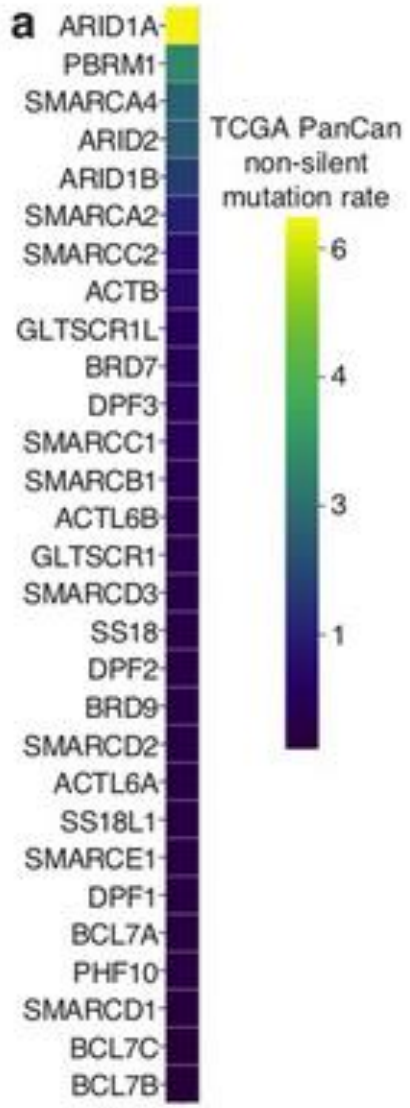
Stacked Bar Charts



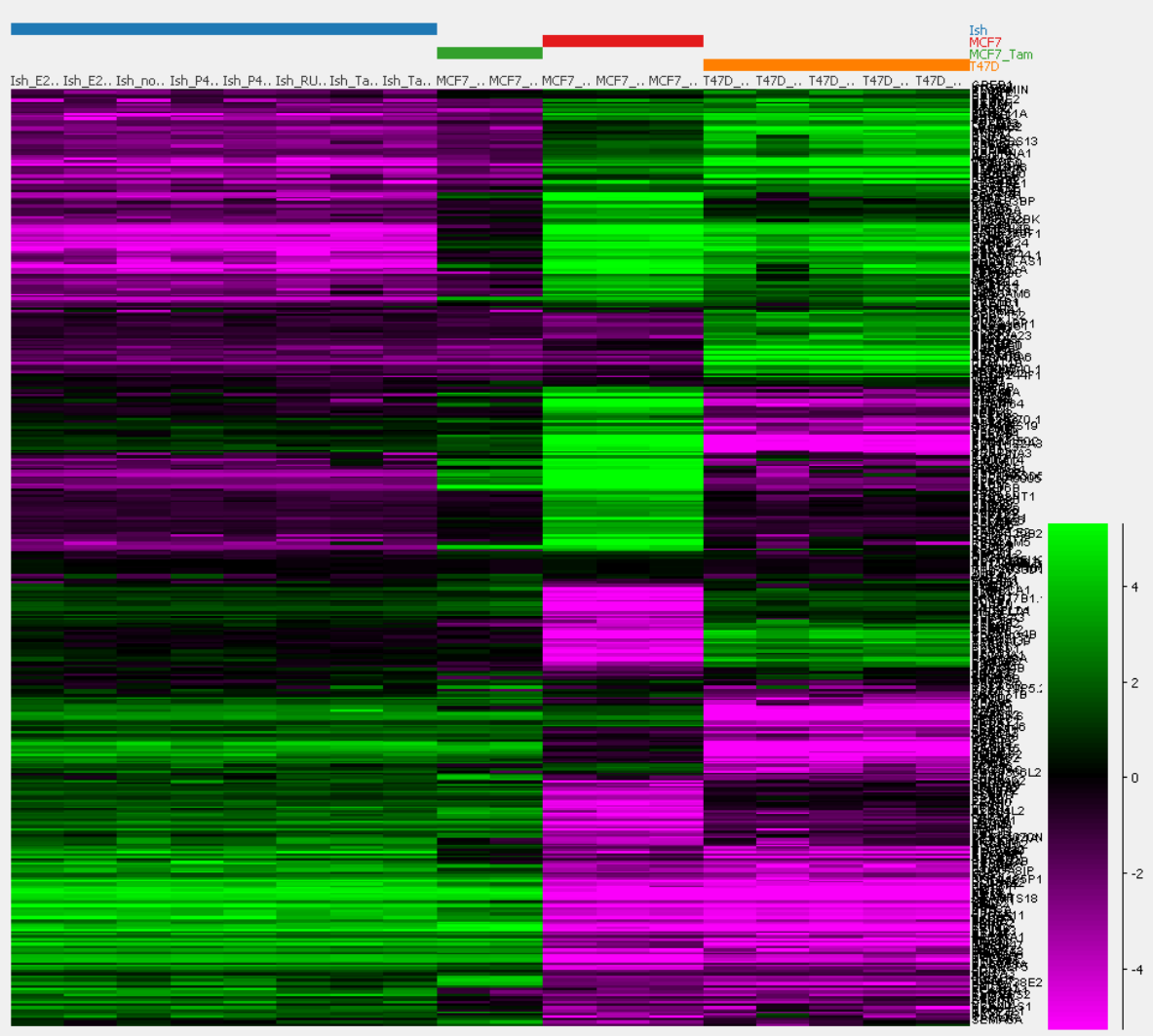


Heatmaps


Ish.. Ish.. Ish.. Ish.. Ish.. Ish.. Ish.. Ish.. MC.. MC.. MC.. MC.. MC.. T4.. T4.. T4.. T4.. T4..
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
4
2
0
-2
-4


a
ARID1A
PBRM1
SMARCA4
ARID2
ARID1B
SMARCA2
SMARCC2
ACTB
GLTSCR1L
BRD7
DPF3
SMARCC1
SMARCB1
ACTL6B
GLTSCR1
SMARCD3
SS18
DPF2
BRD9
SMARCD2
ACTL6A
SS18L1
SMARCE1
DPF1
BCL7A
PHF10
SMARCD1
BCL7C
BCL7B
TCGA PanCan
non-silent
mutation rate
6
4
3
1

Making Heatmaps Effective



- Cluster rows and columns
- Median centre rows
- Diverging symmetrical colour scheme (colourblind friendly)
- Clear annotation

Ethics of data representation

Simon Andrews, Anne Segonds-Pichon

simon.andrews@babraham.ac.uk

Babraham Bioinformatics

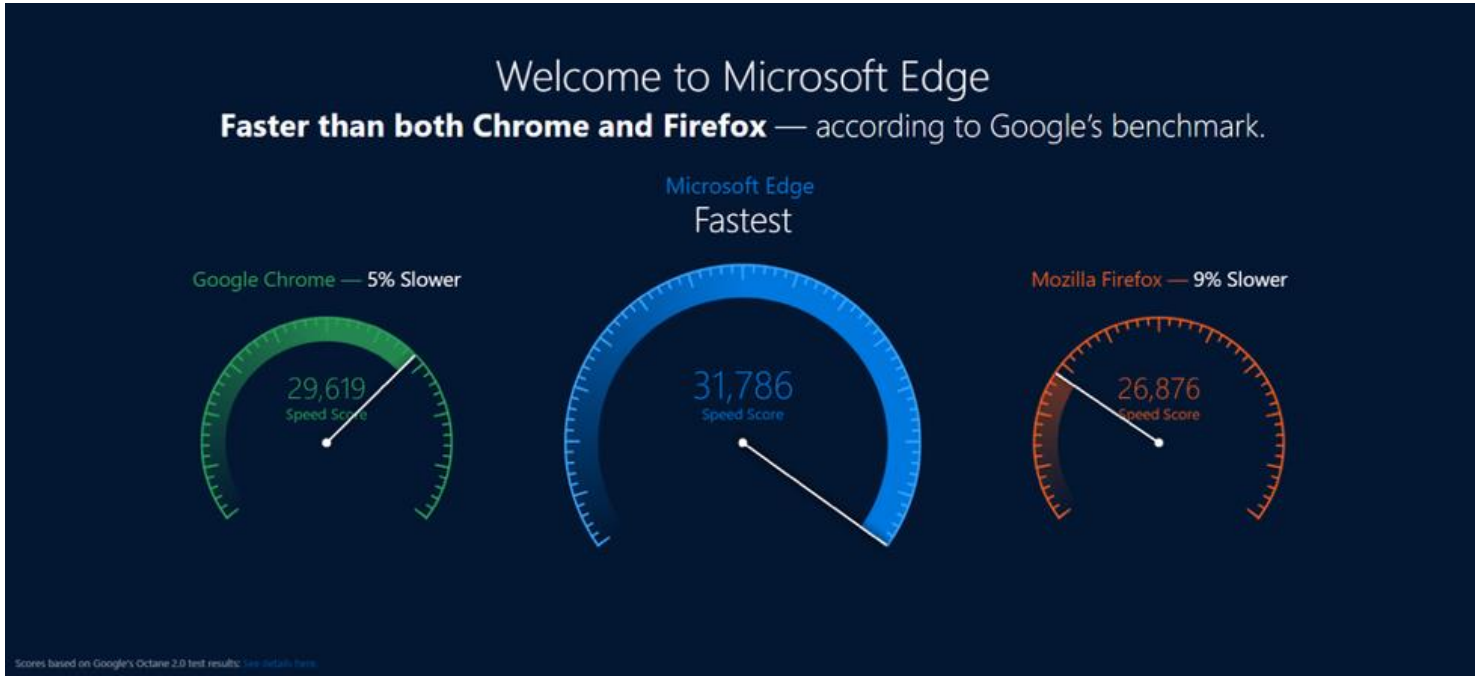
What is an Ethical data visualisation?

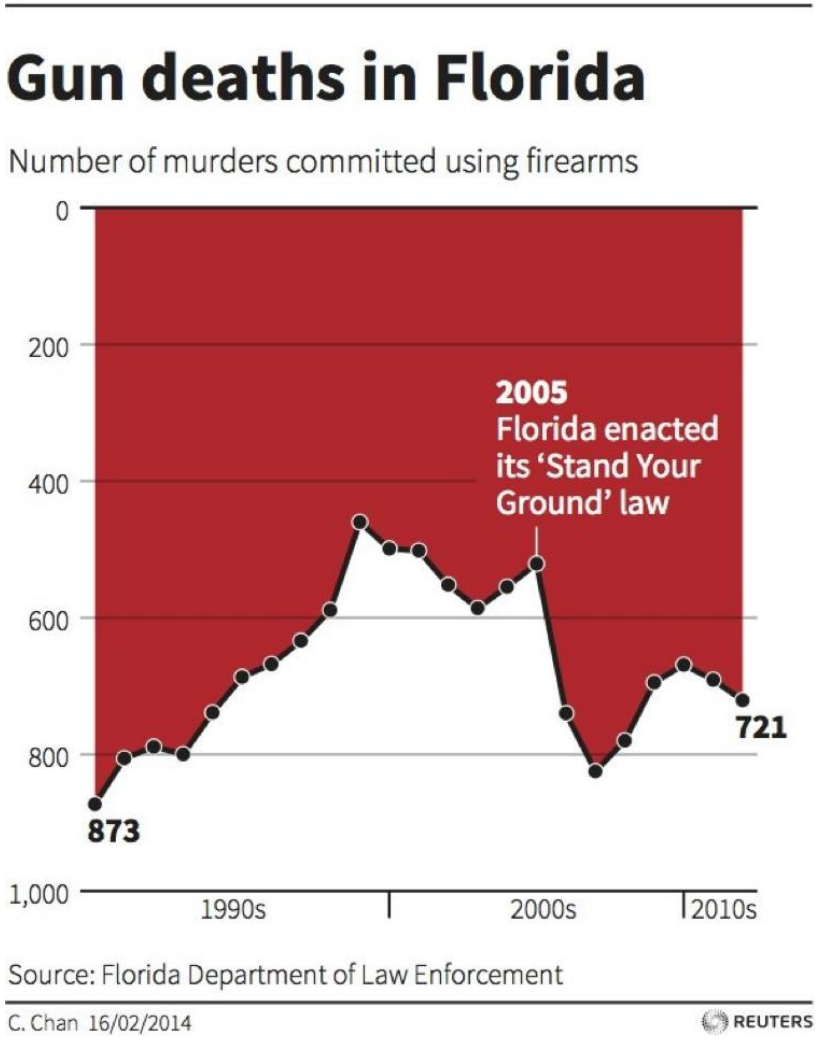
- Different ways of being unethical:
  - not exploring/getting to know the data well enough
  - misusing your chosen graphical representation
  - deliberately showing the data in a misleading manner
  - choosing the ‘most representative’ image/experiment

Is my plot ethical?

Would a reader come to a different conclusion if they could see the details of the data which were omitted from the plot?

Advertising and politics are built on unethical data representation.

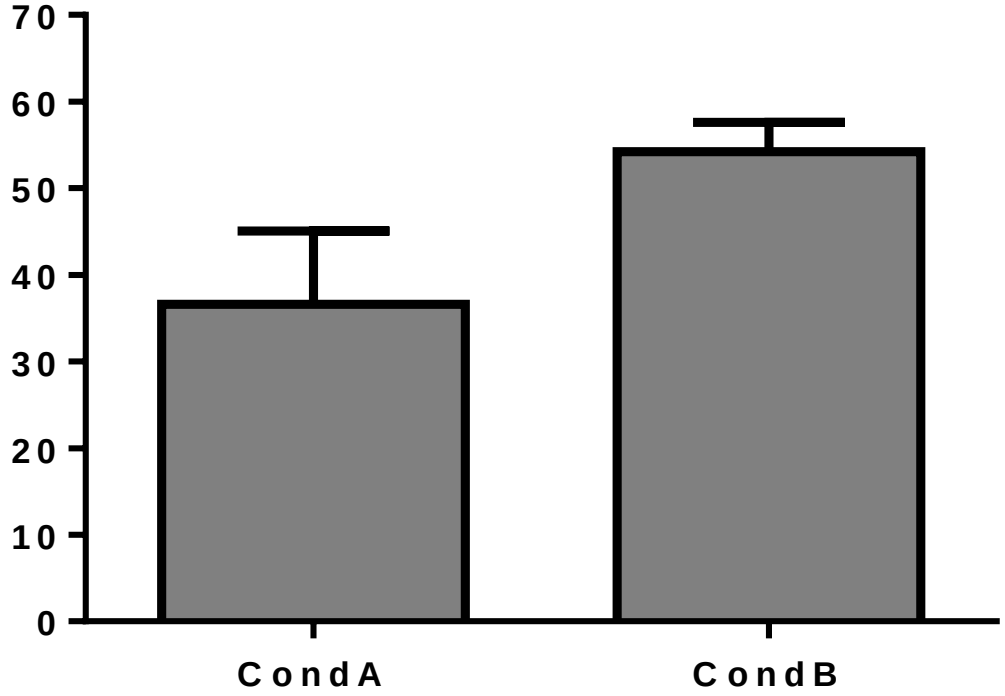


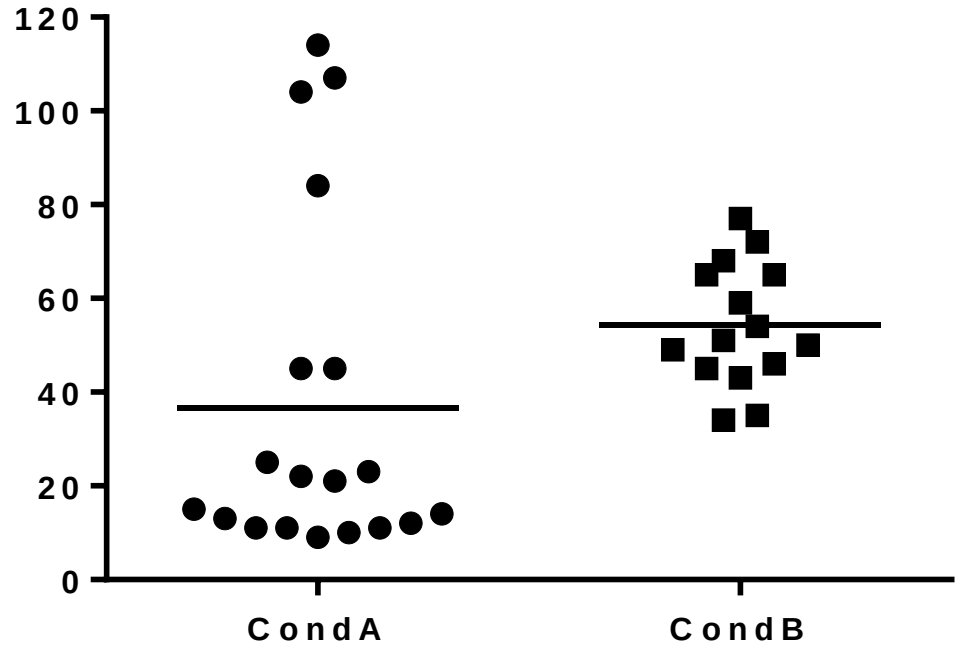


https://venngage.com/blog/misleading-graphs/

Not exploring the data well enough

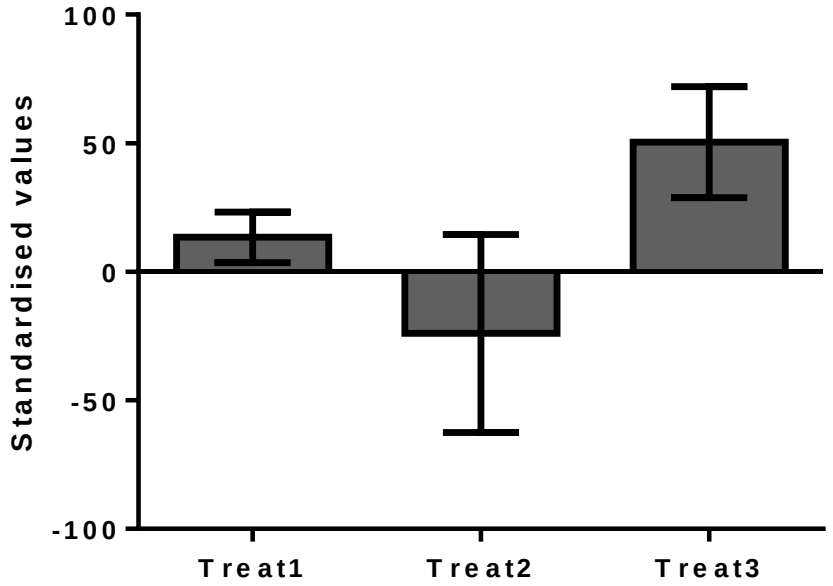
- <u>One experiment</u>: change in the variable of interest between CondA to CondB.
  - Data plotted as a **bar chart**.

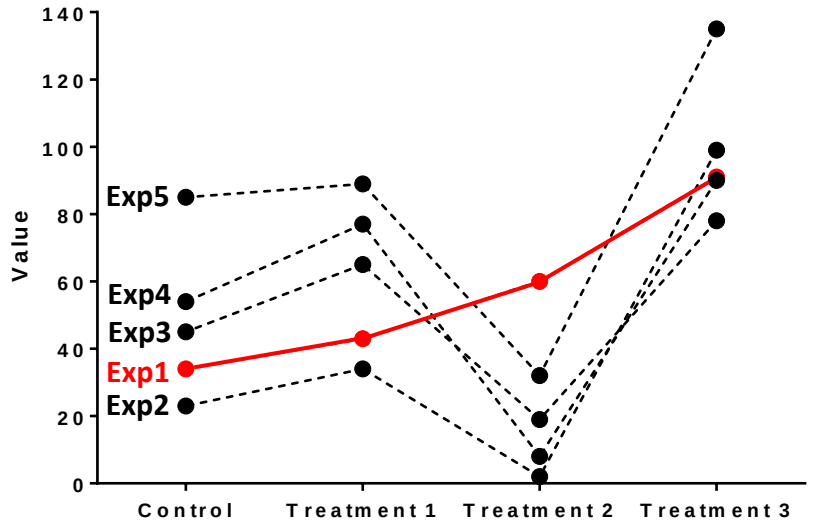




Not exploring the data well enough

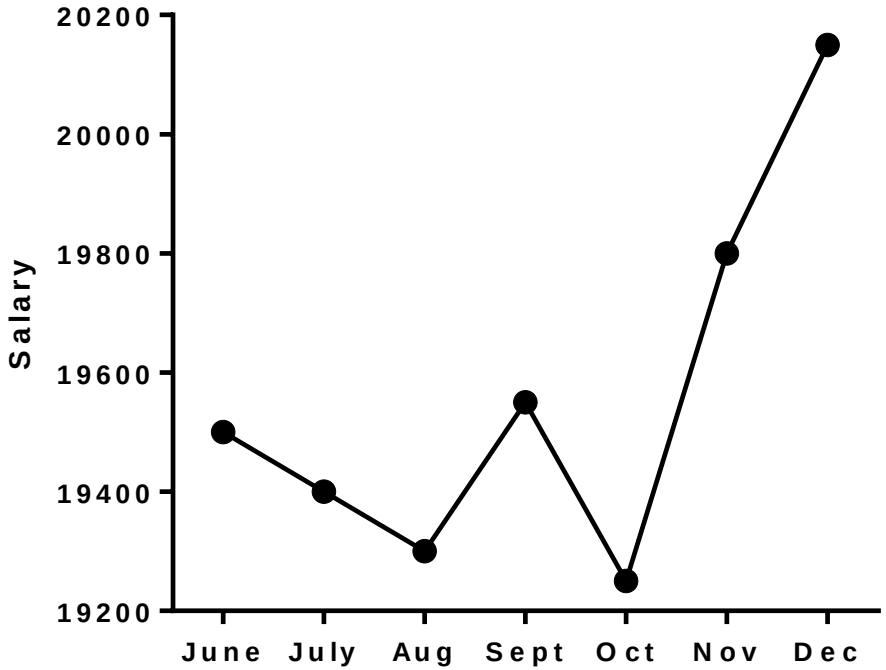
- Five experiments: change in the variable of interest between 3 treatments and a control.
  - Data plotted as a **bar chart**.

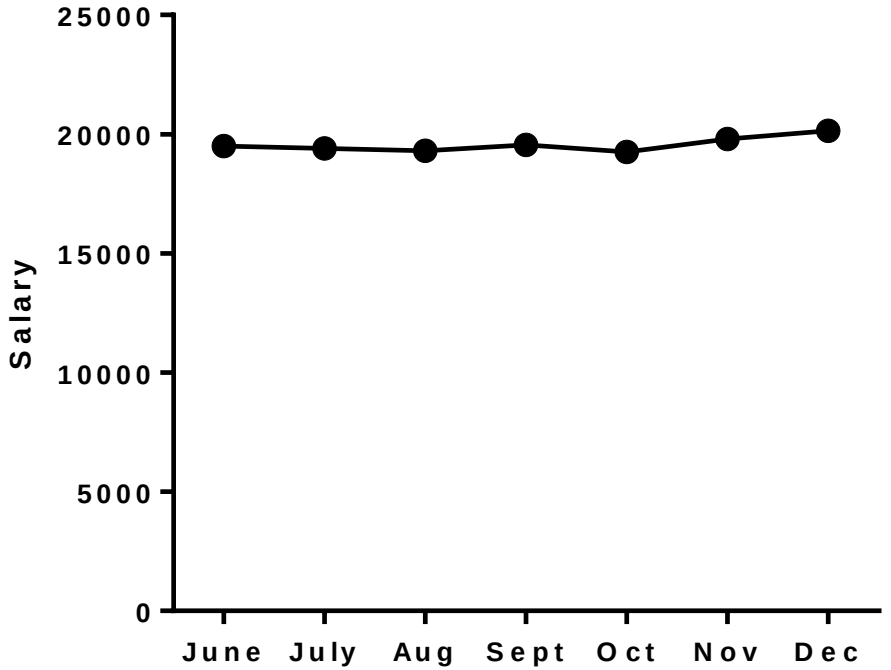




Choosing the wrong axis/scale

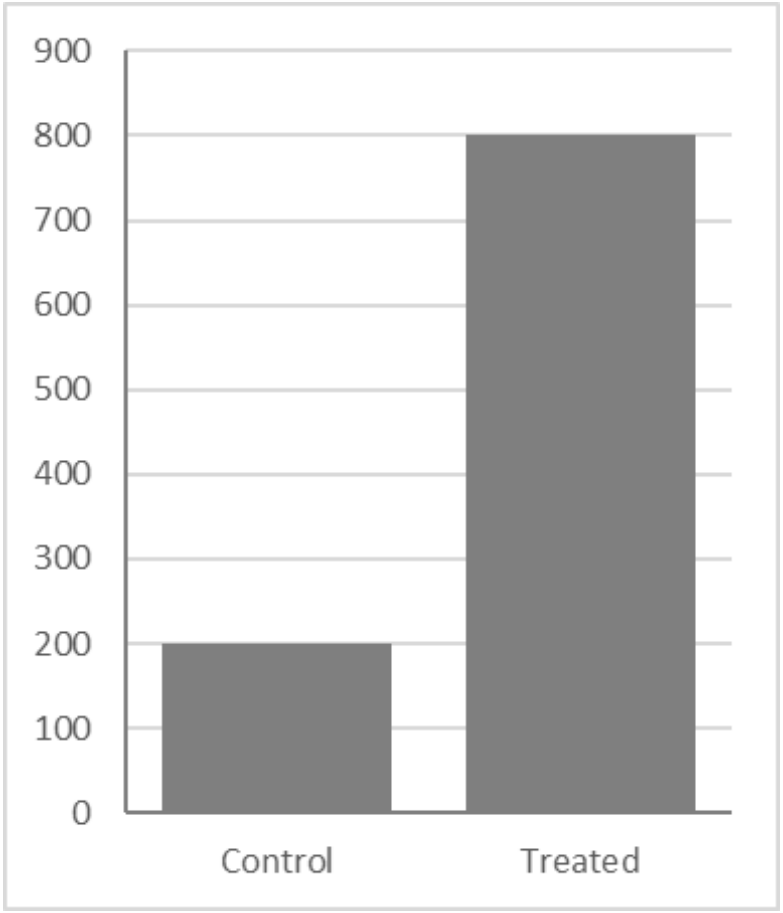
- Example: increase in salaries offered in the last term.

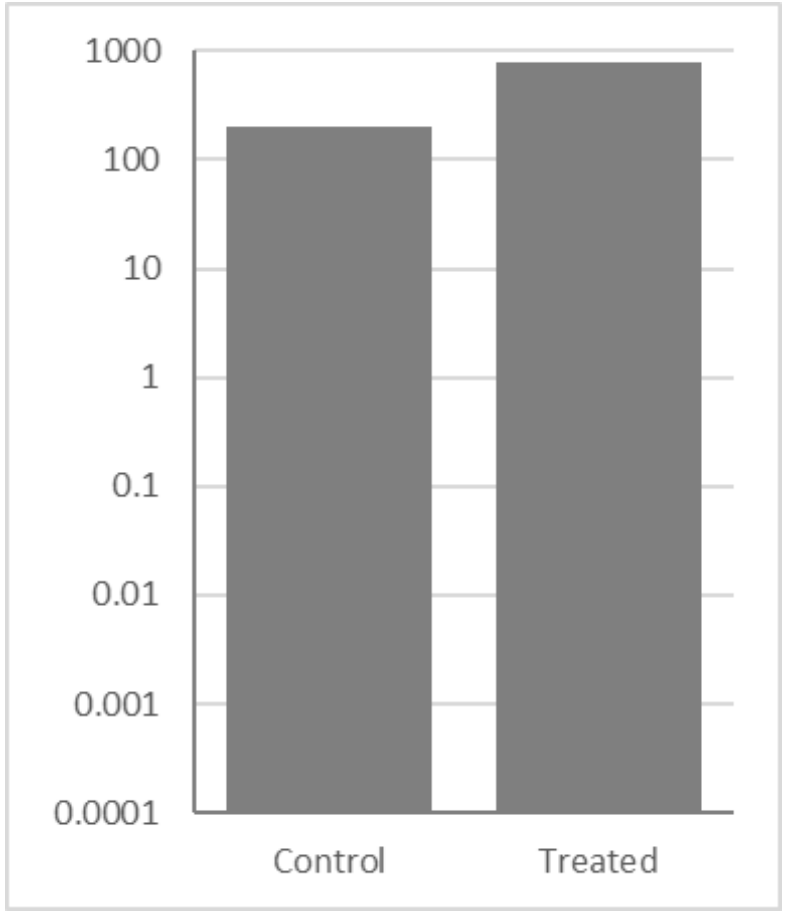




Choosing the y-axis/scale

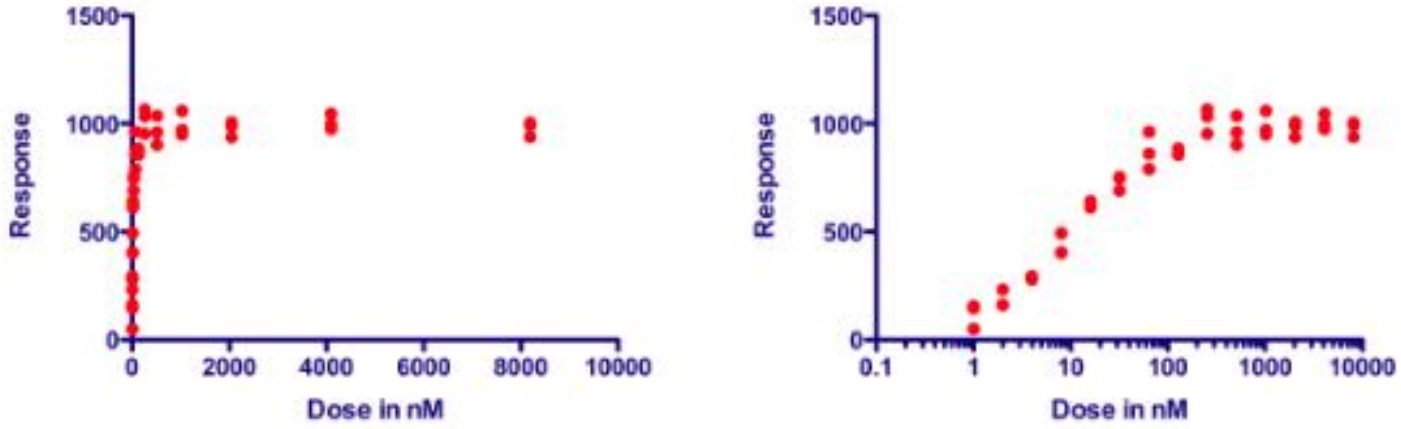
- Inappropriate use of a log scale can artificially minimise differences



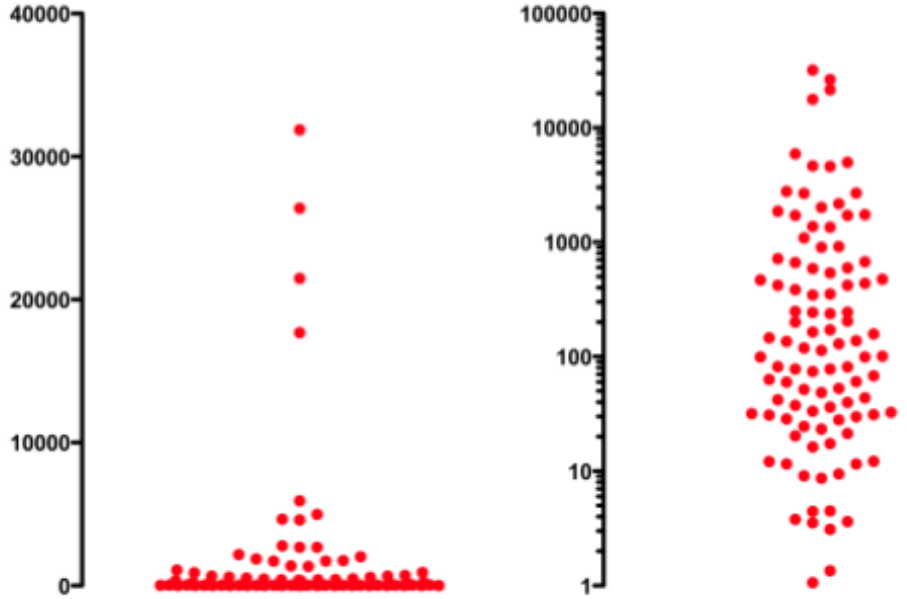


Choosing the y-axis/scale

- **Logarithmic axis** should only be used for:



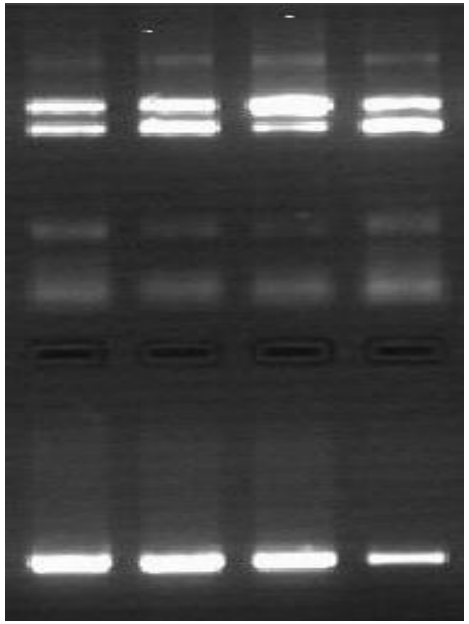
Logarithmically spaced values



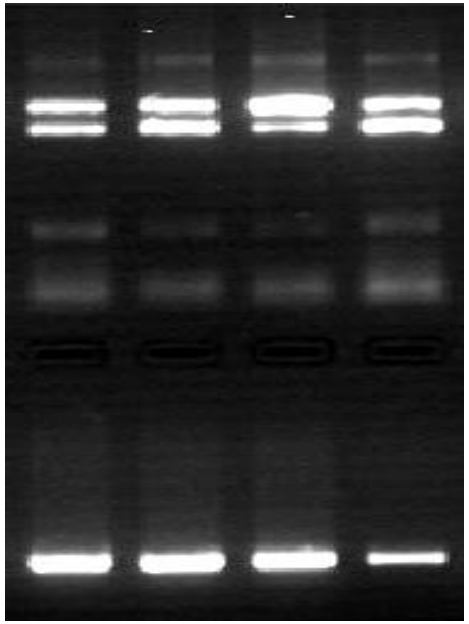
Lognormal data

Image Manipulation

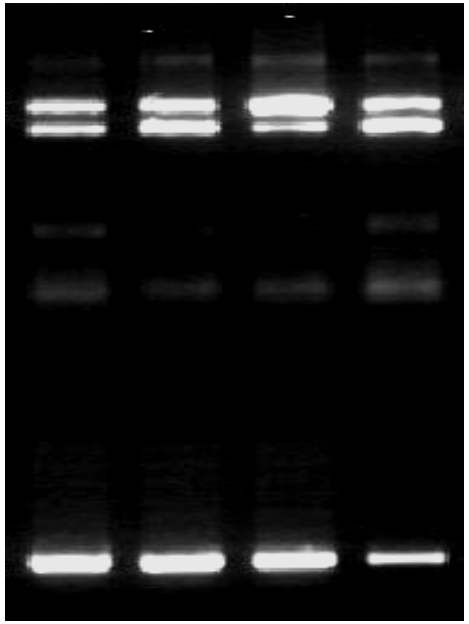
- ‘Playing’ too much with contrast



Original



Brightness and Contrast Adjusted



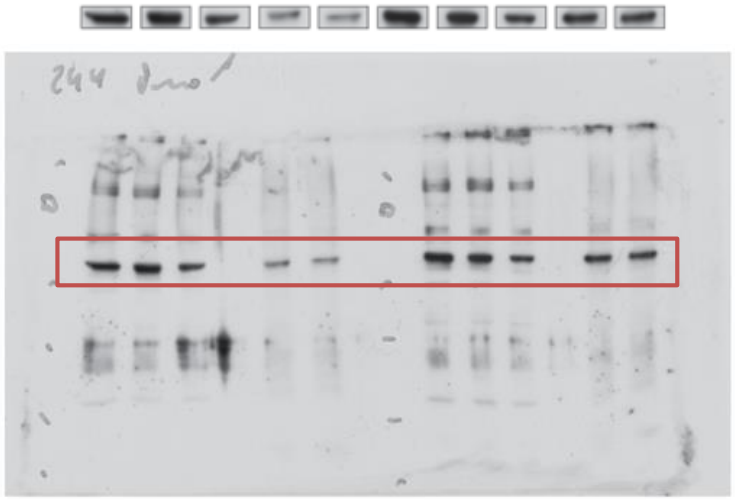
Brightness and Contrast Adjusted Too Much: Oversaturation

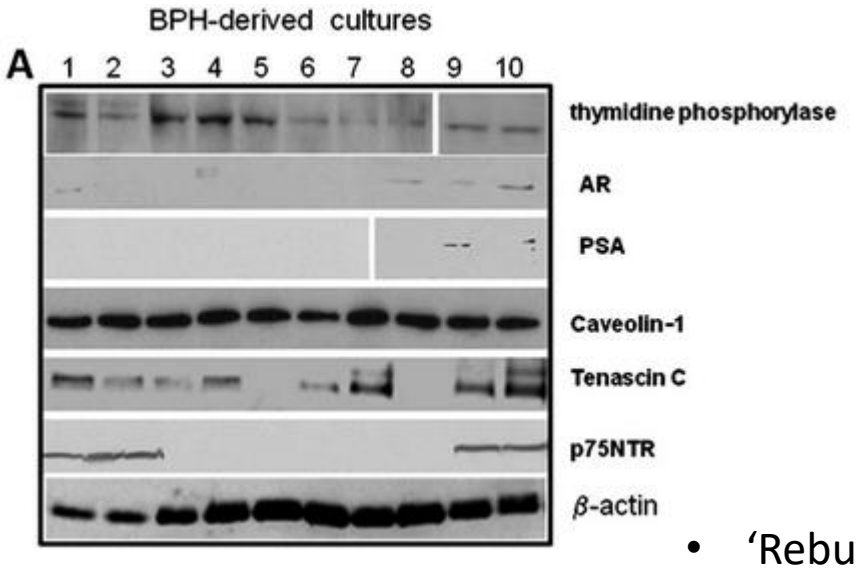
“Adjusting the contrast/brightness of a digital image is common practice and is not considered improper if the adjustment is applied to the whole image.

Adjusting the contrast/brightness of only part of an image is improper, however, and this practice can usually be spotted by someone scrutinizing a file.”

Image Manipulation

- Presenting bands out of context



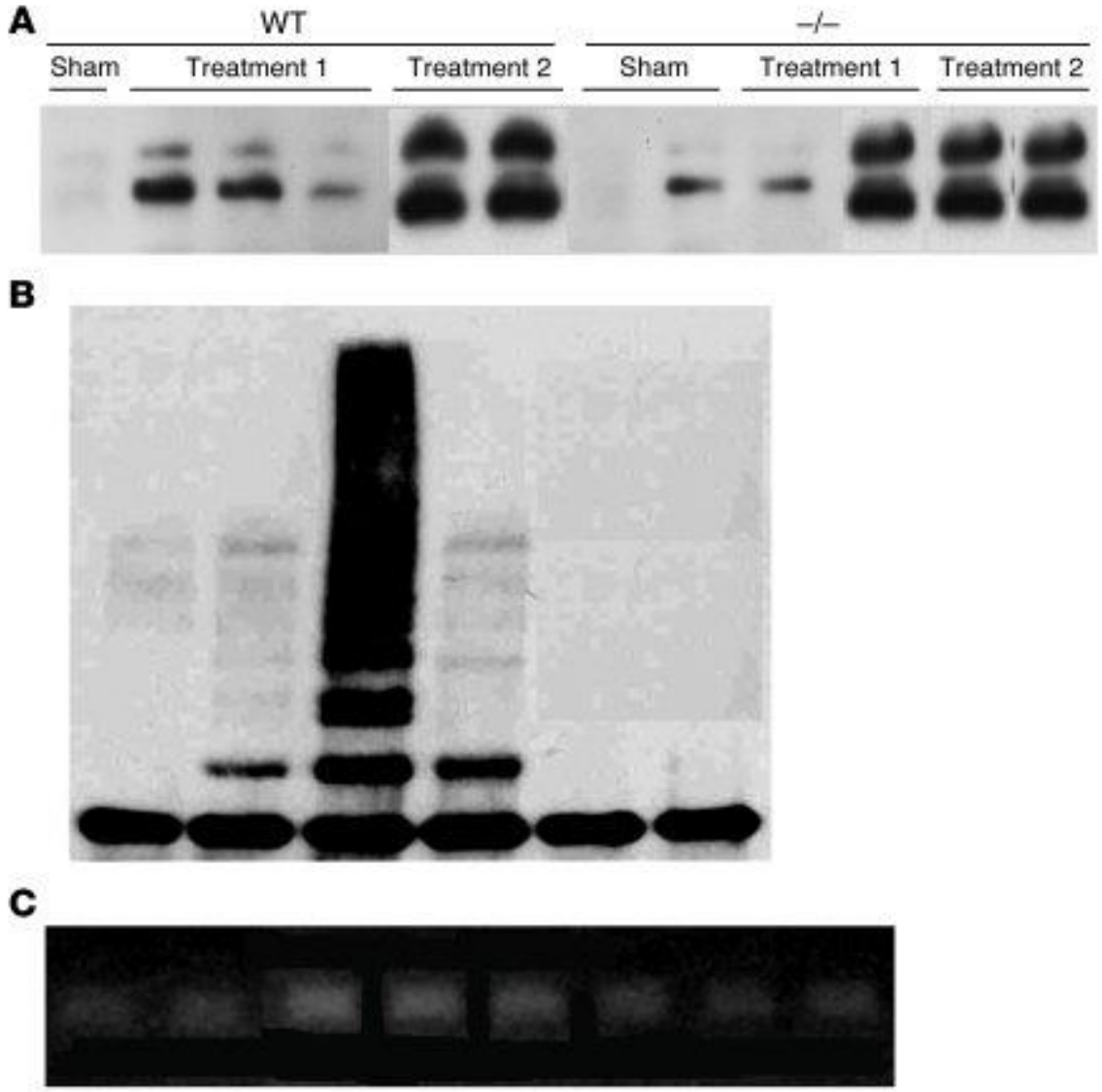


Juxtaposing two lanes that were not next to each other in an original gel is common practice when preparing figures from hard copy photographs of the gel, and is acceptable manipulation if the figure is digital. Taking a band from one digital image and placing it in a lane in another is improper manipulation, which can usually be spotted by someone scrutinizing a file.

- 'Rebuilding' a gel from several cuts

Image Manipulation can be detected

Image Data Integrity



10.1172/JCI28824

Is my plot ethical?

Would a reader come to a different conclusion if they could see the details of the data which were omitted from the plot?

Practical Design Theory

Boo Virk
Simon Andrews

simon.andrews@babraham.ac.uk

Babraham Bioinformatics

Why does good design matter?

- Good design makes a great first impression
- Good design makes for effective communication
- Good design keeps the reader engaged

Art Palvanov (http://www.palvanov.com/)

Planning

- Always look at the guidelines for the journal you're submitting to
  - https://www.sciencemag.org/authors/instructions-preparing-initial-manuscript
  - https://www.nature.com/nature/for-authors/formatting-guide
  - https://www.cell.com/figureguidelines
- Huge variation in the amount of detail they provide
- Getting things right from the start saves huge amounts of time

General Figure Guidelines

- Use distinct colors with comparable visibility and consider colorblind individuals by avoiding the use of red and green for contrast. Recoloring primary data, such as fluorescence images, to color-safe combinations such as green and magenta, turquoise and red, yellow and blue or other accessible color palettes is strongly encouraged. Use of the rainbow color scale should be avoided.

- Use solid color for filling objects and avoid hatch patterns.

- Avoid background shading.

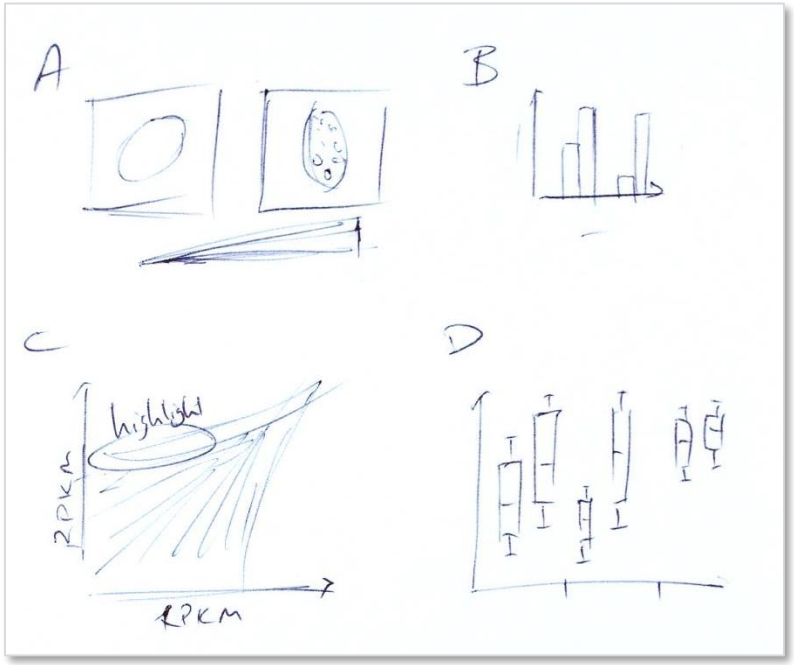
- Figures divided into parts should be labeled with a lower-case, boldface 'a', 'b', etc in the top left-hand corner. Labeling of axes, keys and so on should be in 'sentence case' (first word capitalized only) with no full stop. Units must have a space between the number and the unit, and follow the nomenclature common to your field.

- Commas should be used to separate thousands.

- Unusual units or abbreviations should be spelled out in full, or defined in the legend.

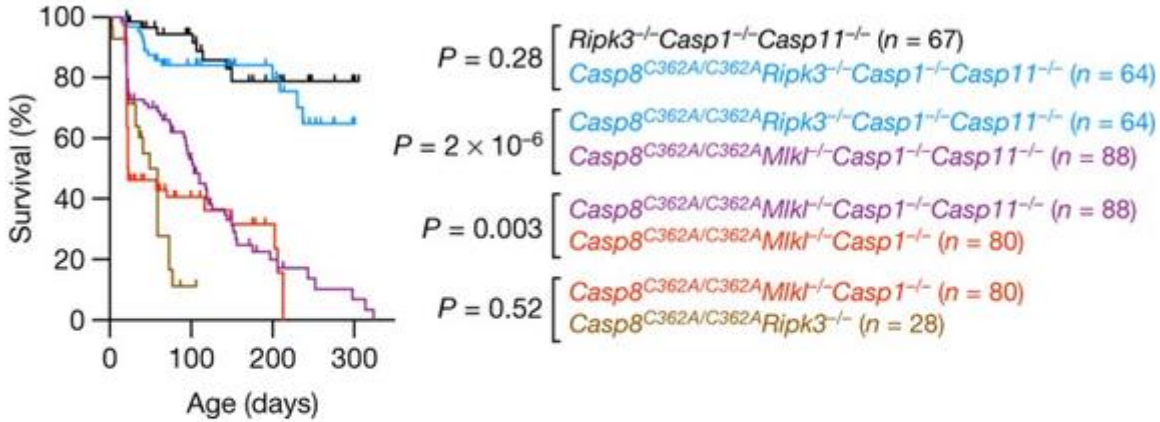
https://mts-ncomms.nature.com/cgi-bin/main.plex?form_type=display_auth_instructions

Plan out your panels



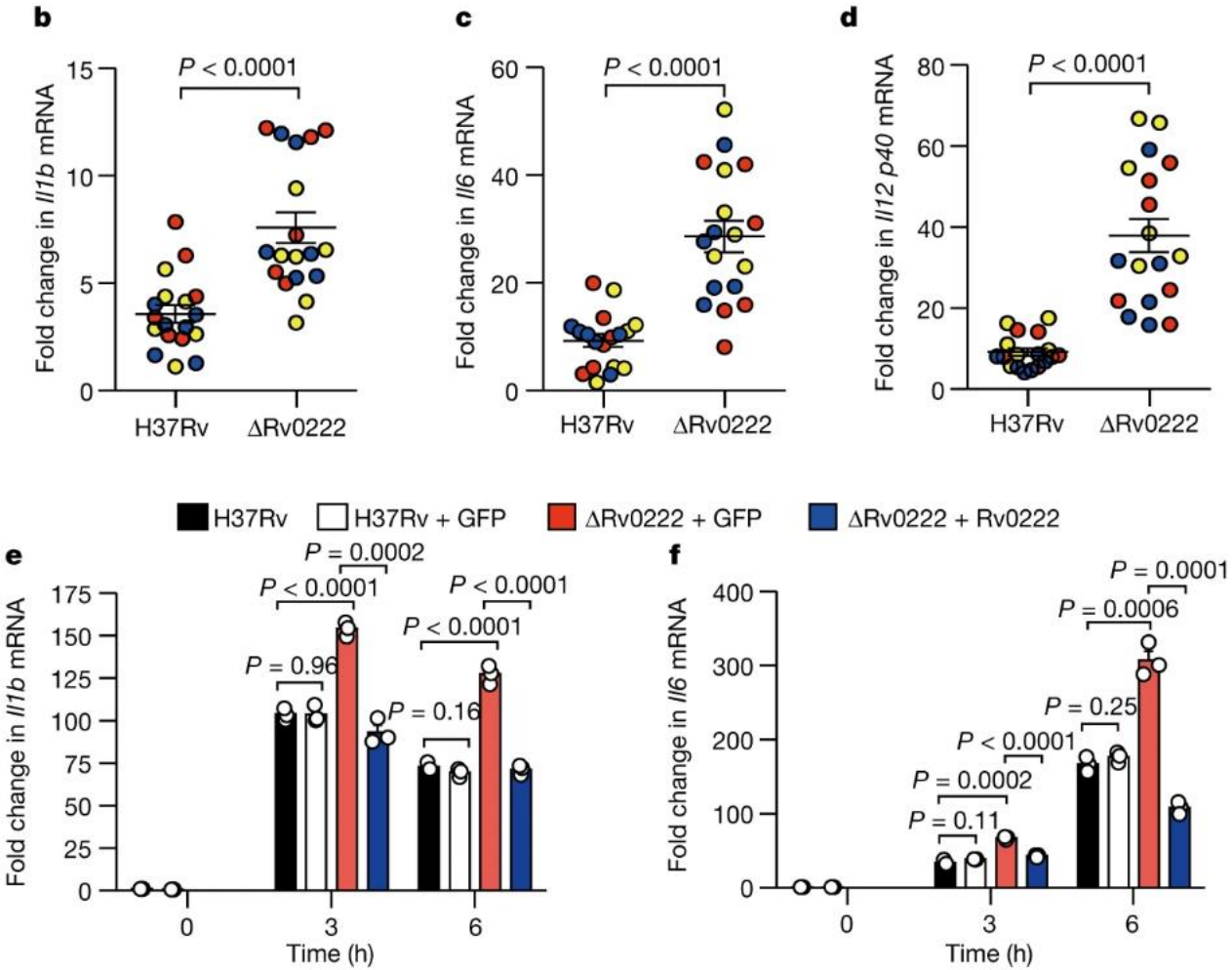
- Plan your panels before starting to draw final figures
- Plan to be consistent
  - Multiple figures of the same type
  - Common colour/shape schemes
  - Common fonts and sizing
  - Common abbreviations and units
  - Common naming of samples / conditions

Fig. 2: Caspase-1, caspase-11 and RIPK3 promote lethality in *Casp8*$^{C362A/C362A}$*Mlkl*$^{-/-}$ mice.

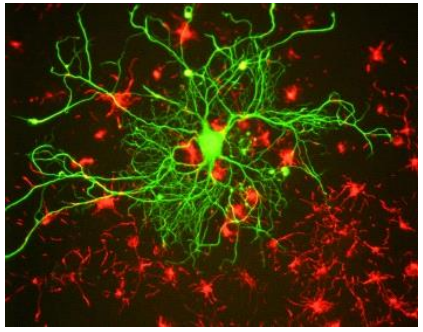


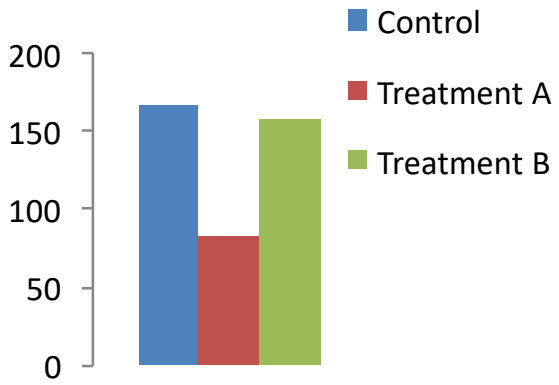
Kaplan–Meier curves of mouse survival. *P* values were calculated by two-sided Gehan–Breslow–Wilcoxon test. The number of mice differs from the list in Table 1, as some of the mice in the graph had a *Casp8*$^{C362A/C362A}$ parent.

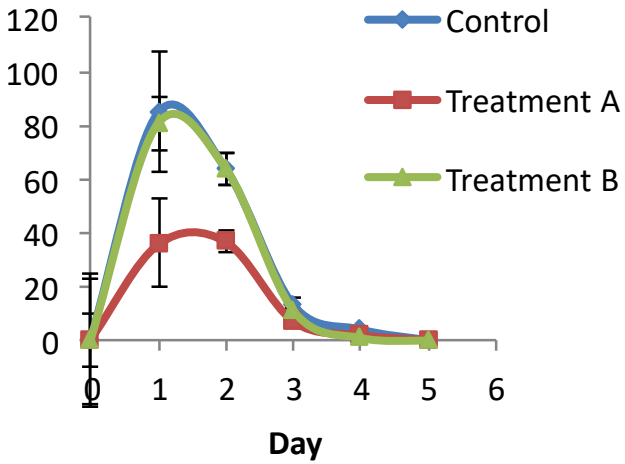
Source data.

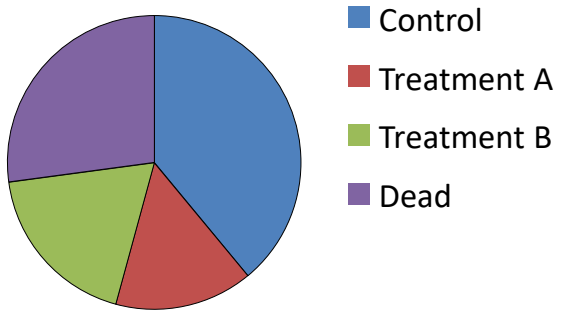


Alignment: We are sensitive to aligned edges, even when they are separated

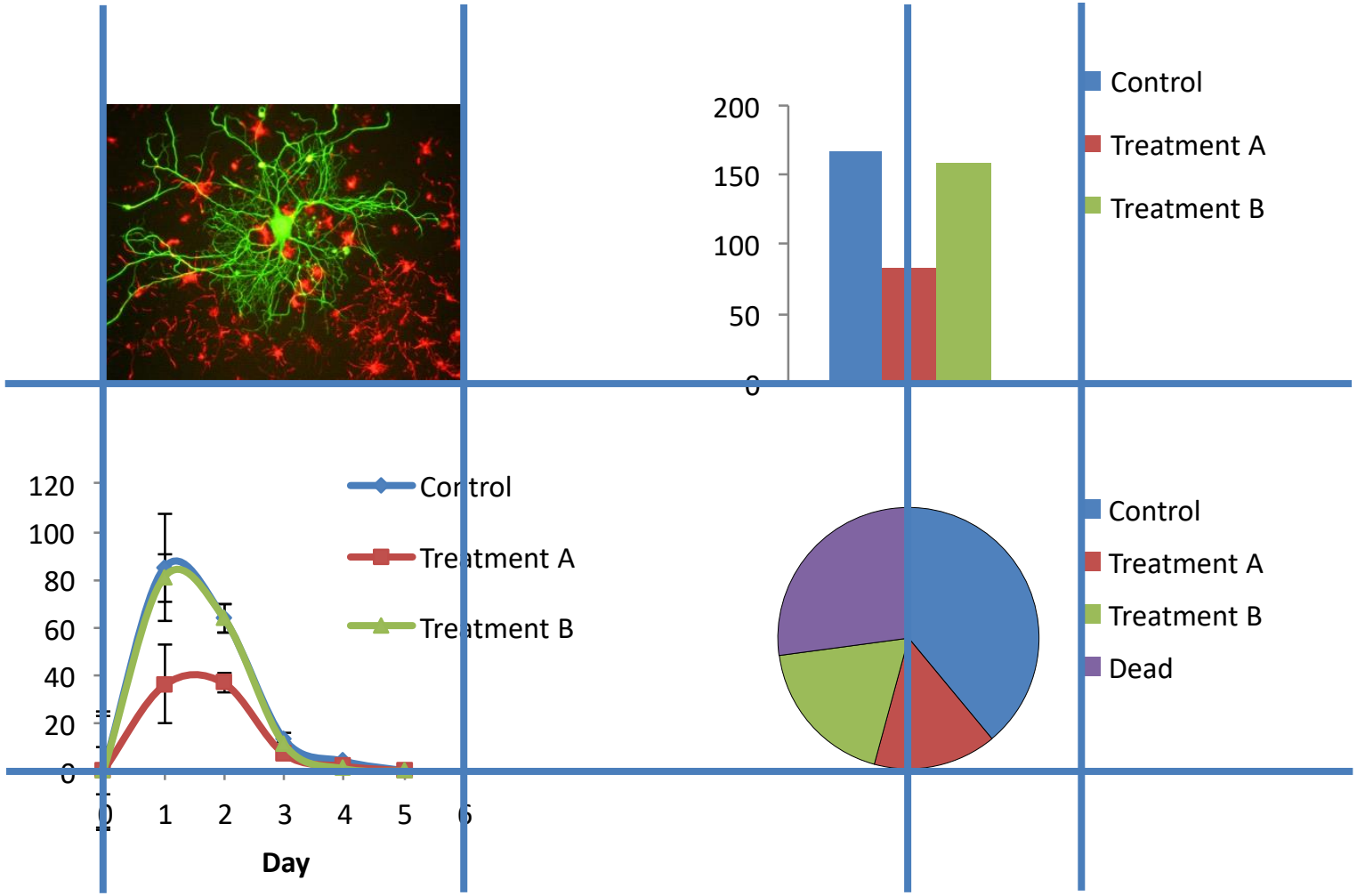




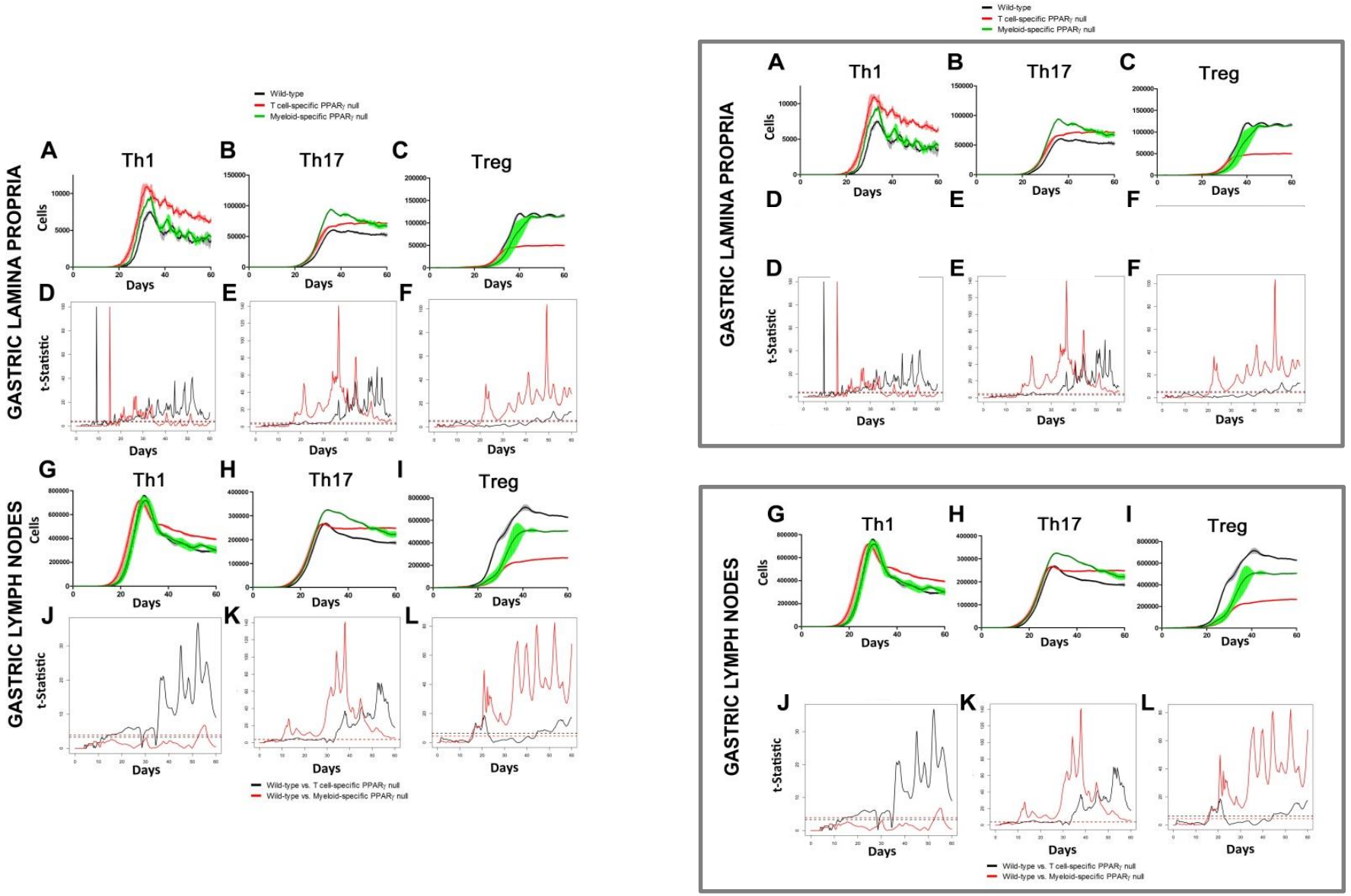




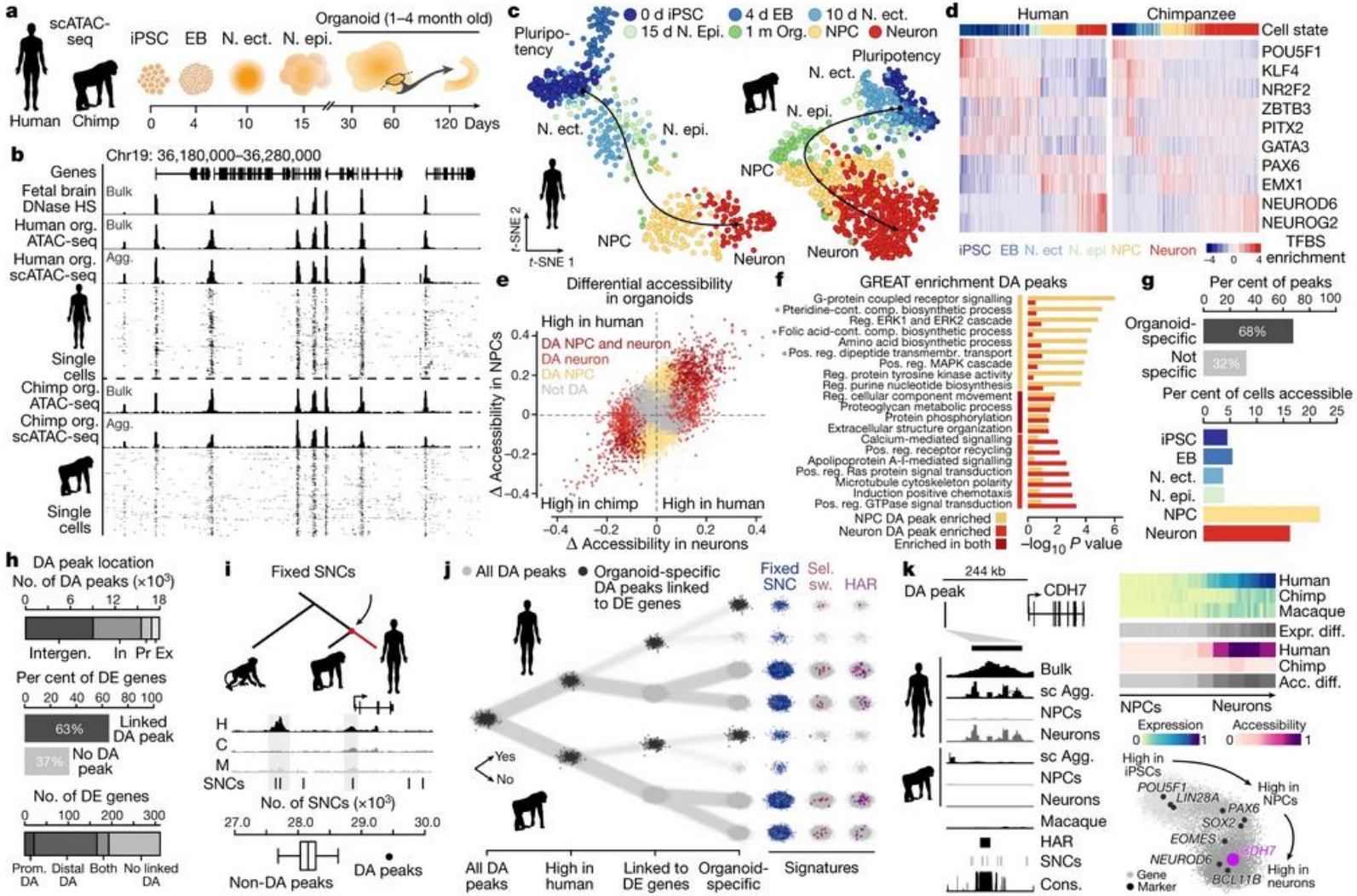
Use a grid to help align disparate parts of a figure



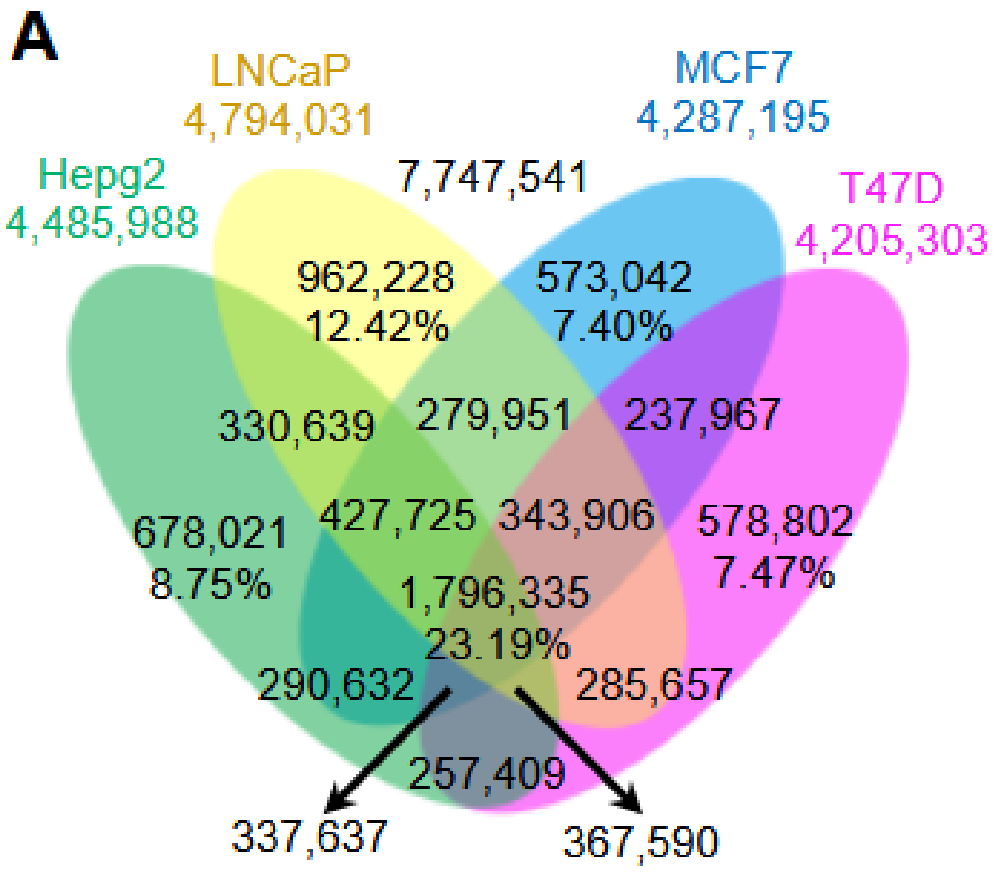
Don't make figures too crowded

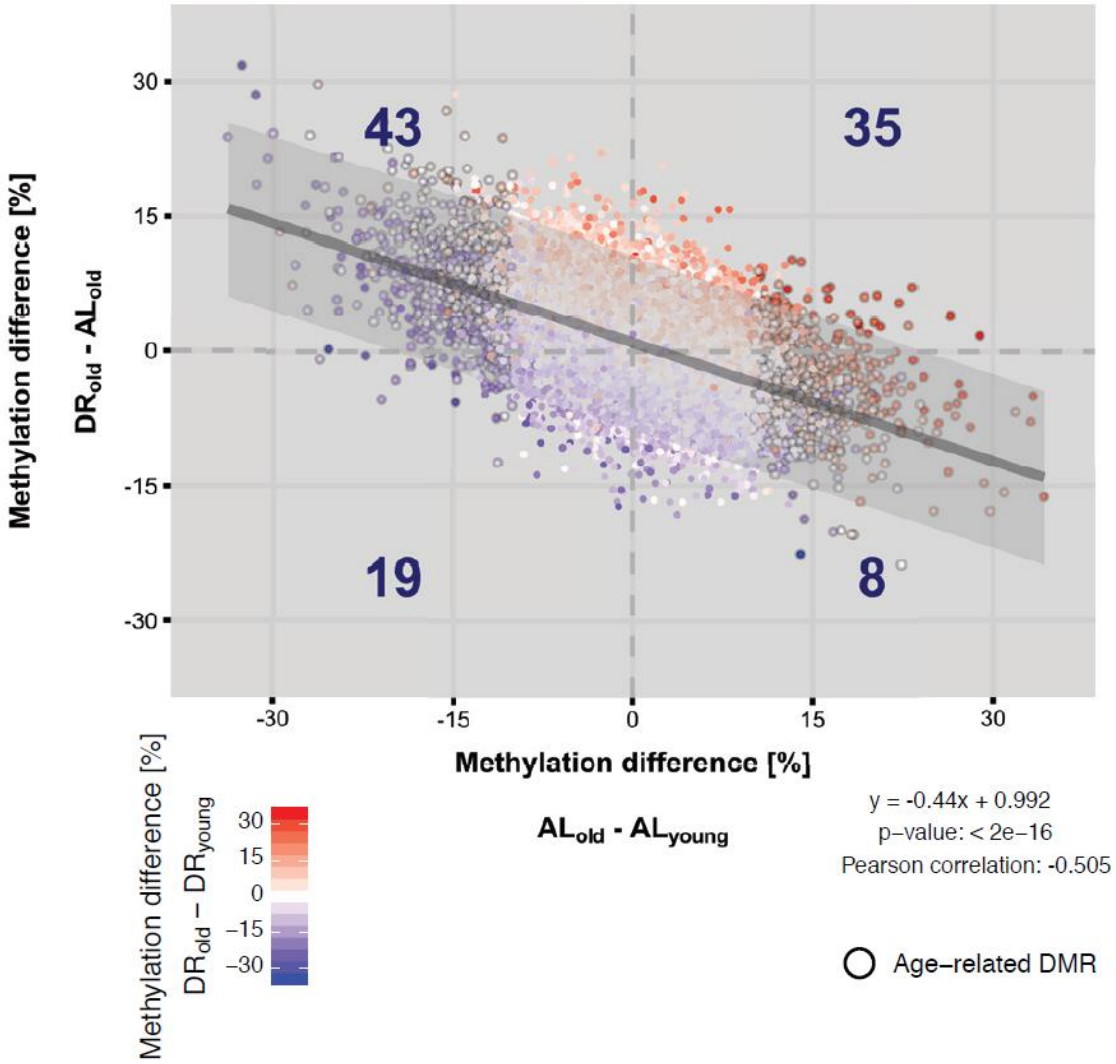


Don't make figures too crowded

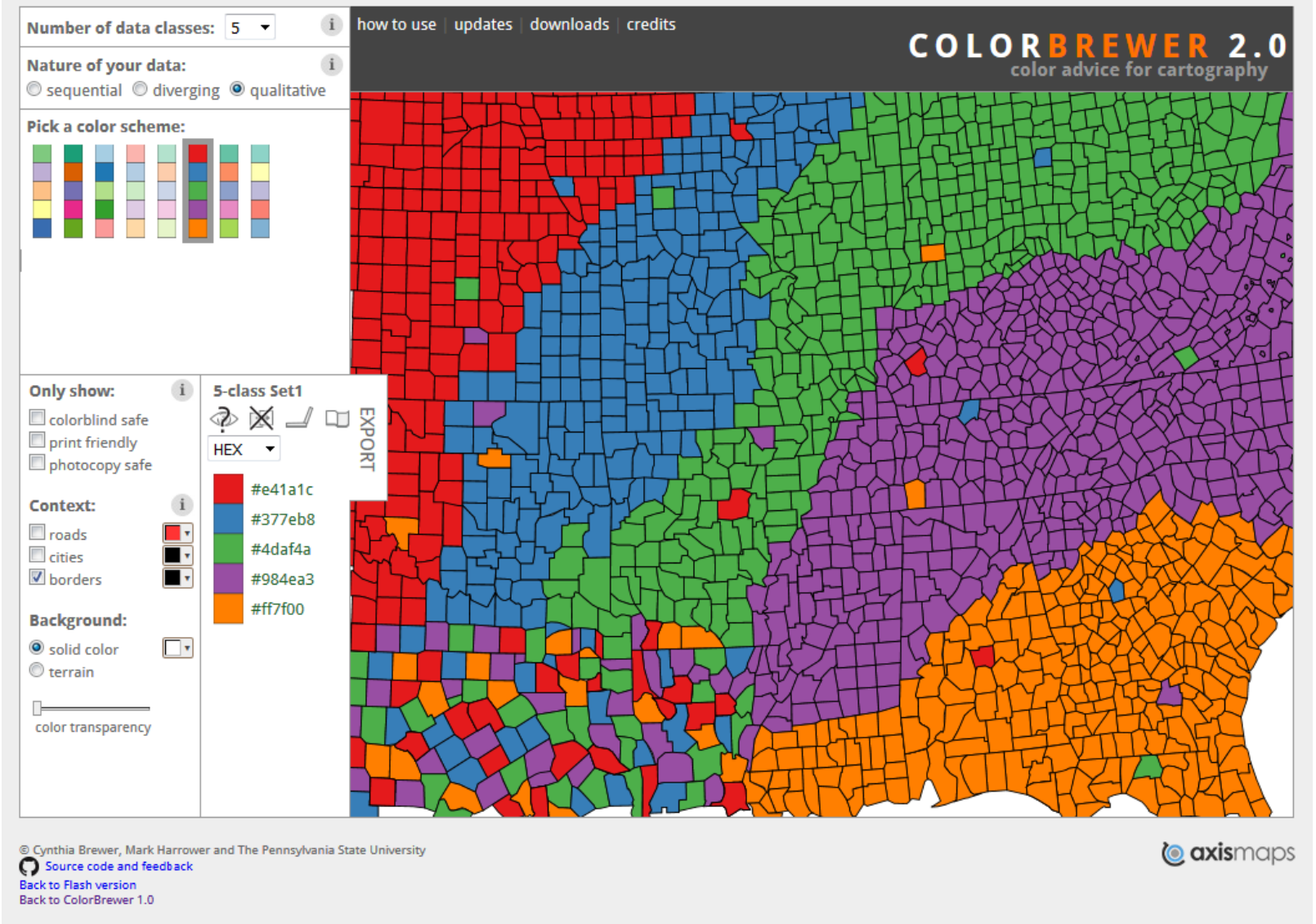


Don't cram too much information onto one figure





Don’t invent your own colour schemes



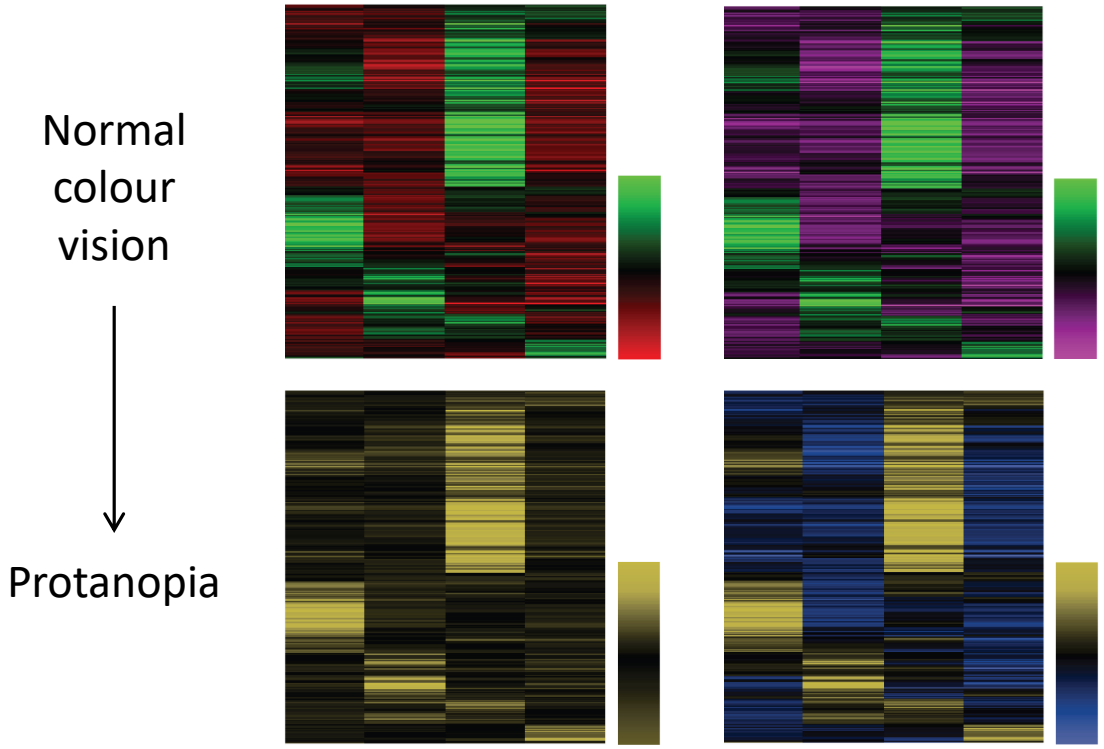
Colorbrewer2.org

If possible try to consider colour blind readers

- Affects 1:12 men and 1:200 women worldwide
- ***“If a submitted manuscript happens to go to three male reviewers of Northern European descent, the chance that at least one will be colour blind is 22 percent.”***

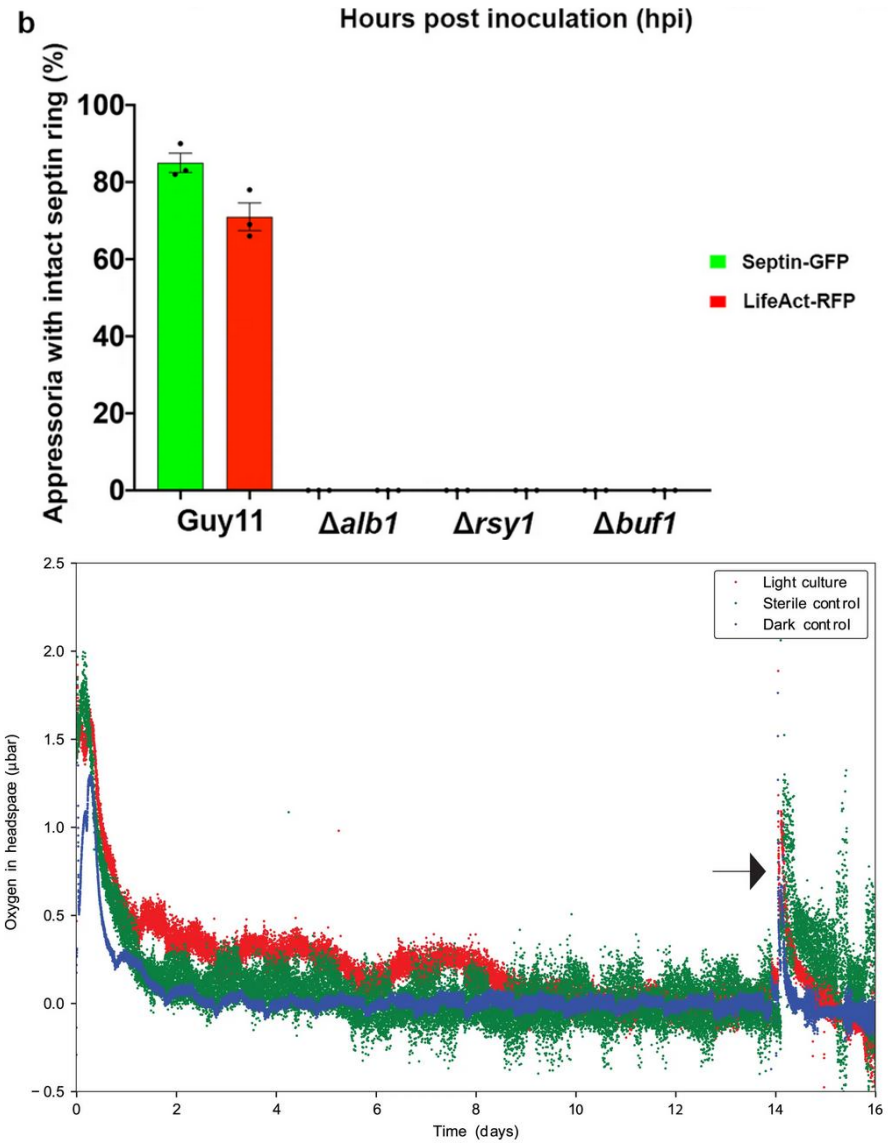
See how well your figure works for colour blind people

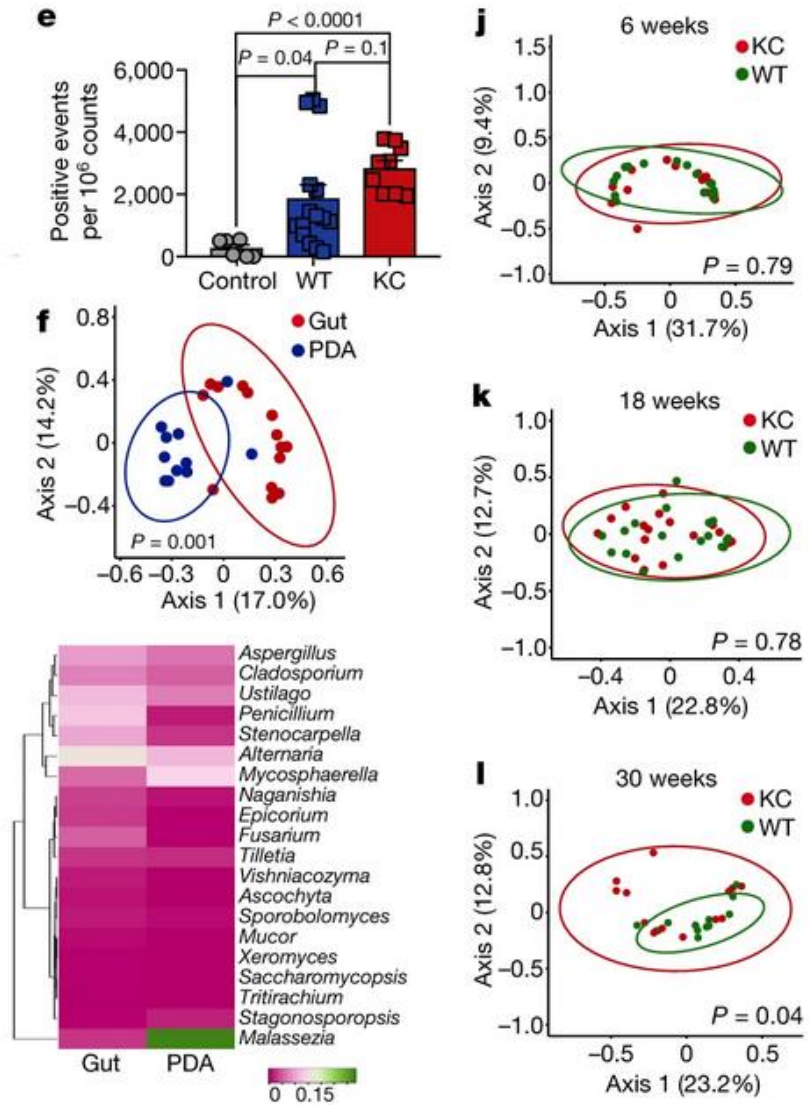
- Gradients are easy to change
- Categorical colours are very limited
- Basic interpretability in black and white is ideal



http://www.color-blindness.com/coblis-color-blindness-simulator/

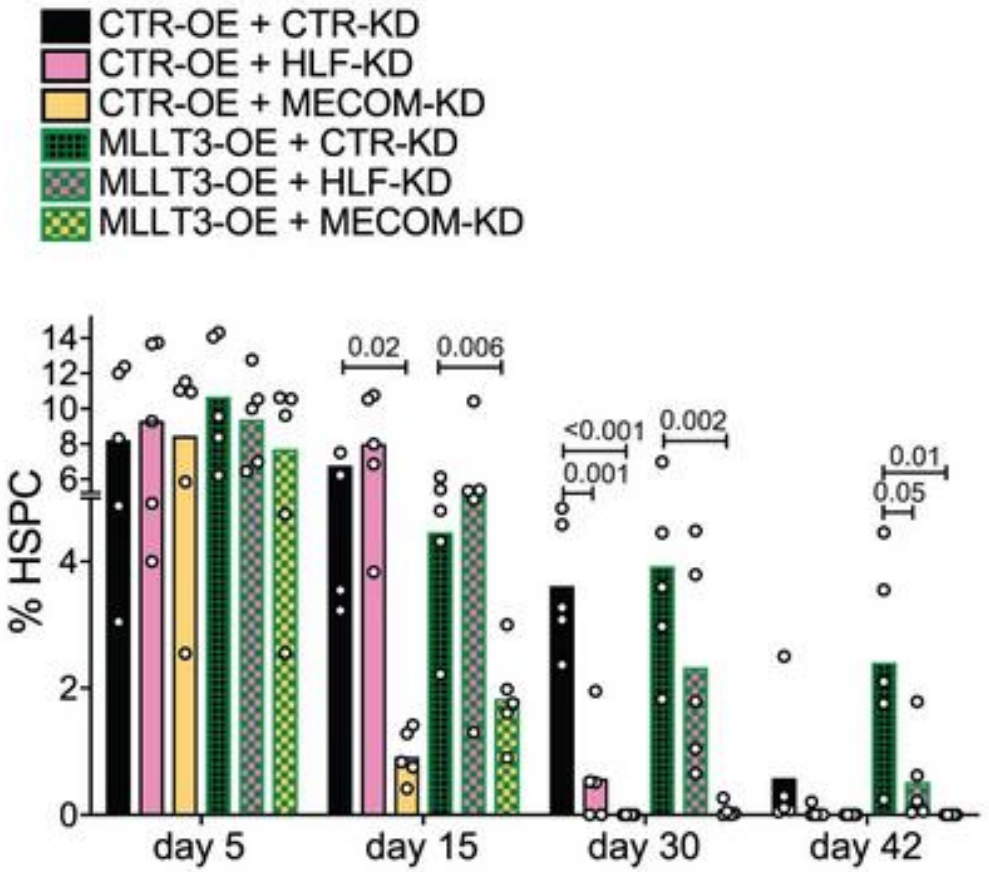
Try to consider colour blind readers





Only use plain colours as fills

- Use a standard colour scheme
- Optimise for colour blind people if possible
- Keep colours plain



When overlaying information, make sure you have sufficient contrast

Poor contrast

Good contrast

Poor contrast

Good contrast

Vibrating colour

Add overlays to increase contrast

Keep text and fonts simple

- All fonts for figures should use sans serif fonts

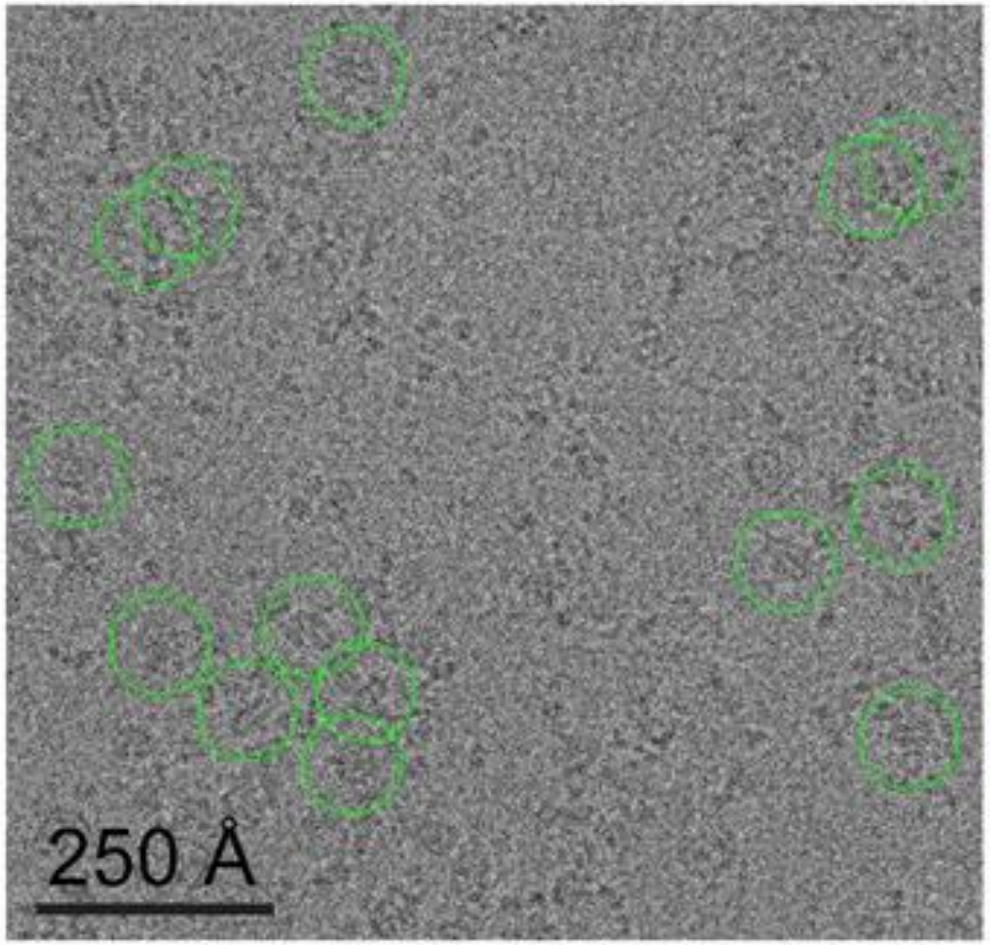
sans-serif serif

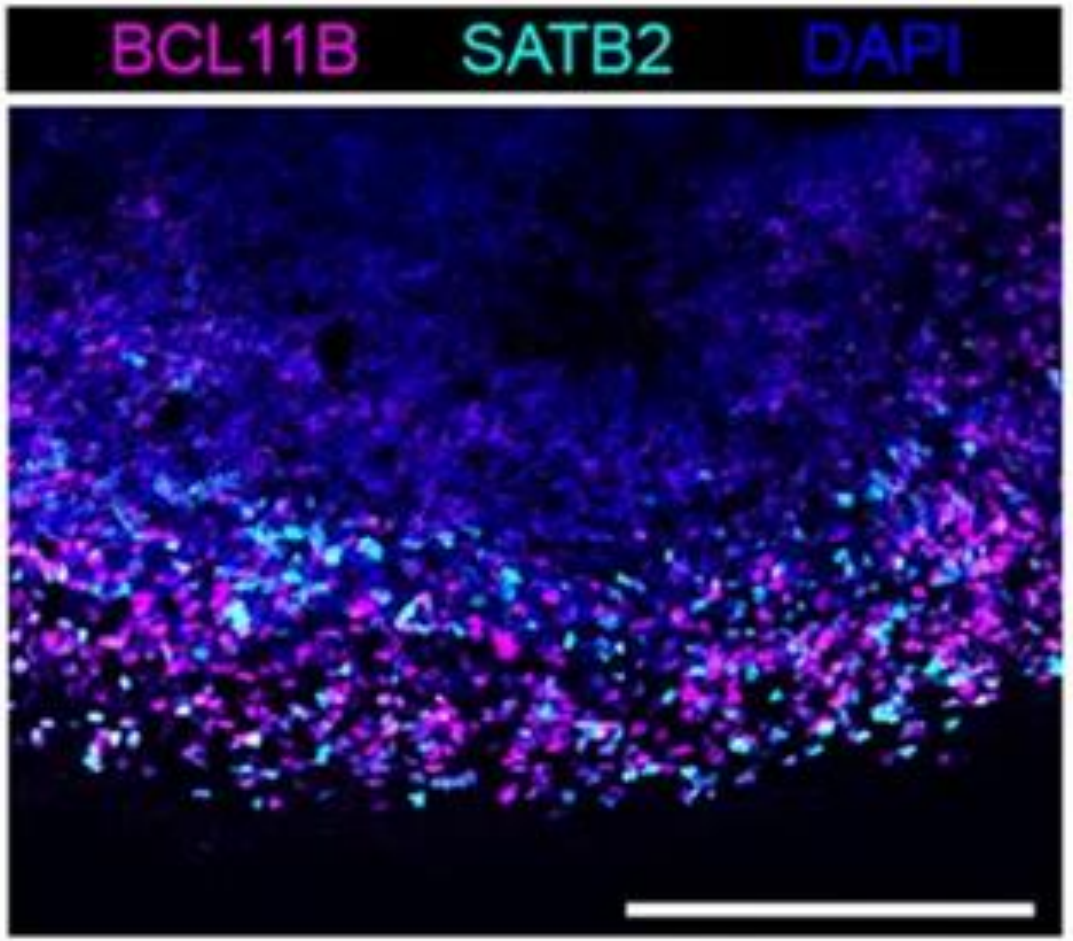
- All text in figures should be black or white*

* Some journals insist on coloured text. They're wrong, but you can't fight the system

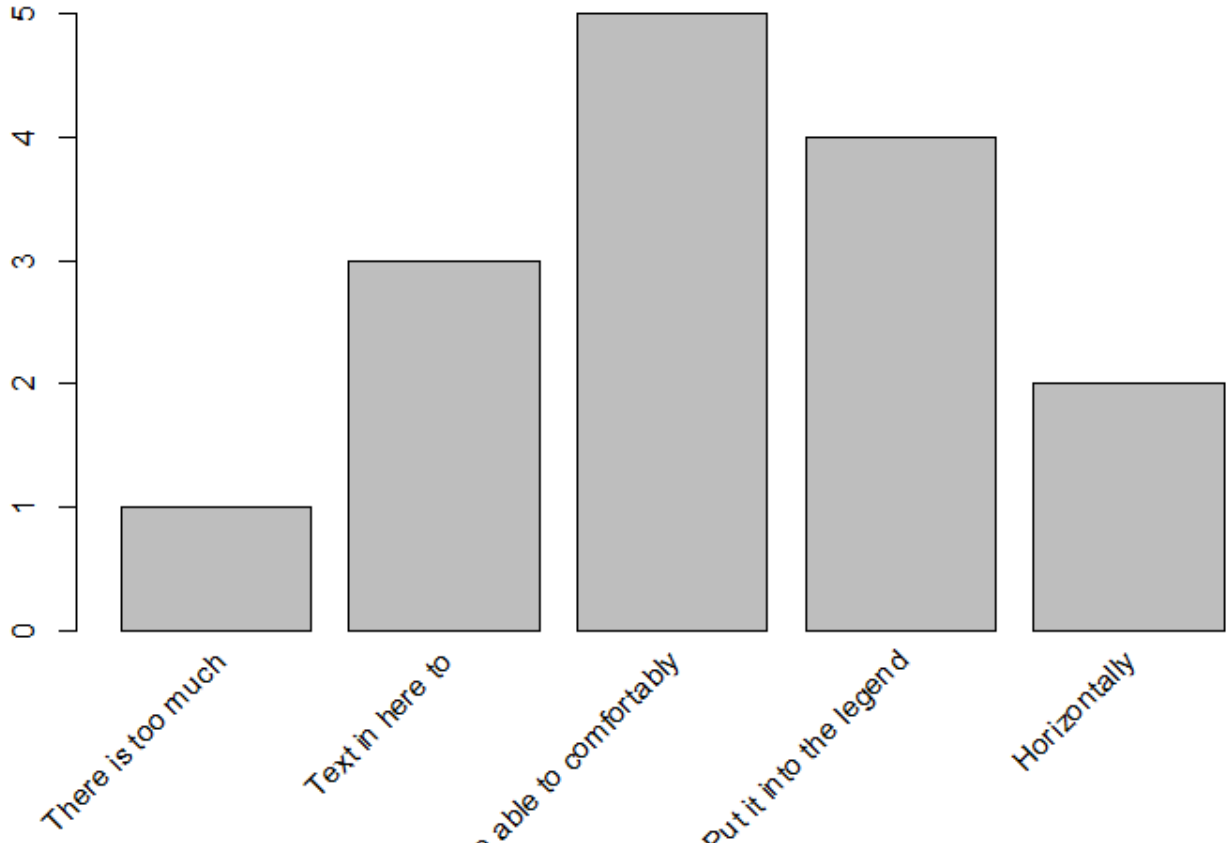
Contrast and text

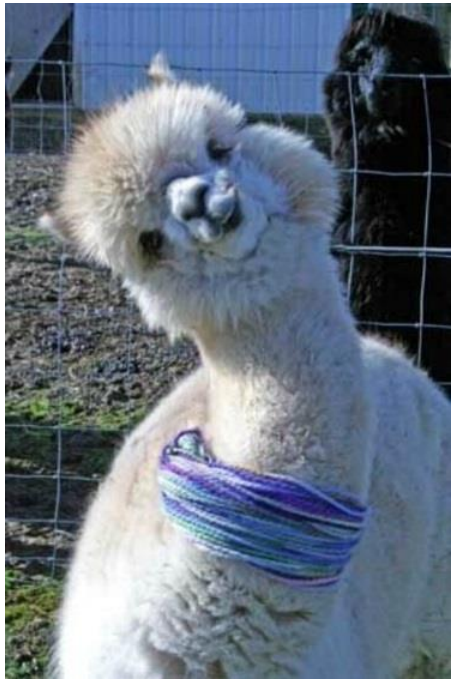
a



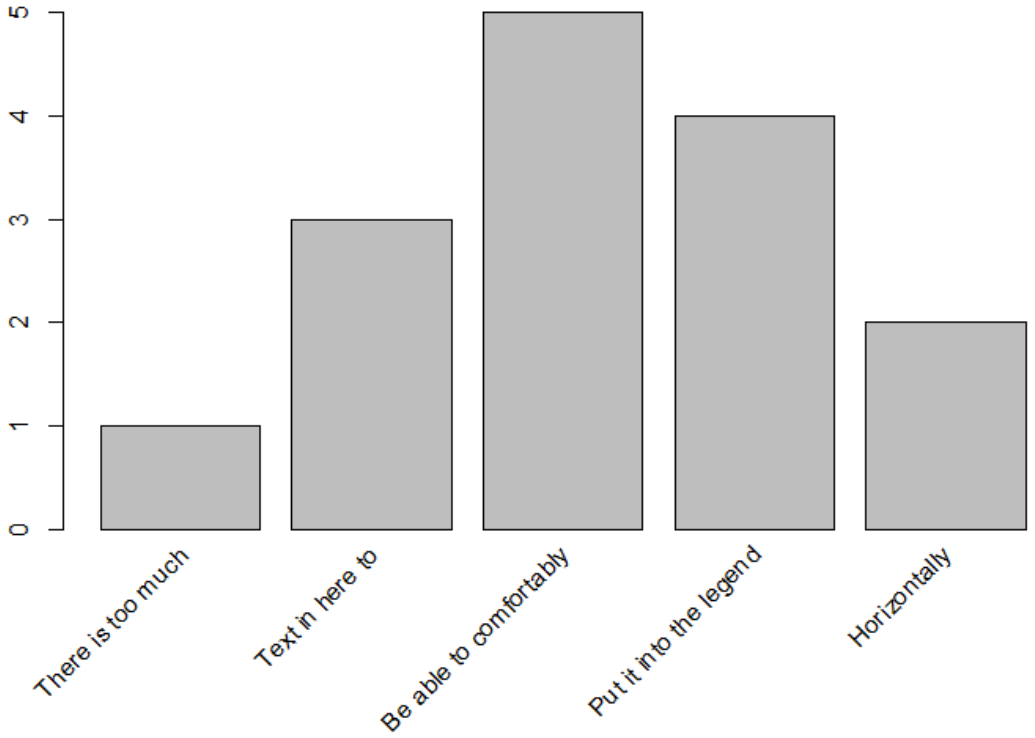


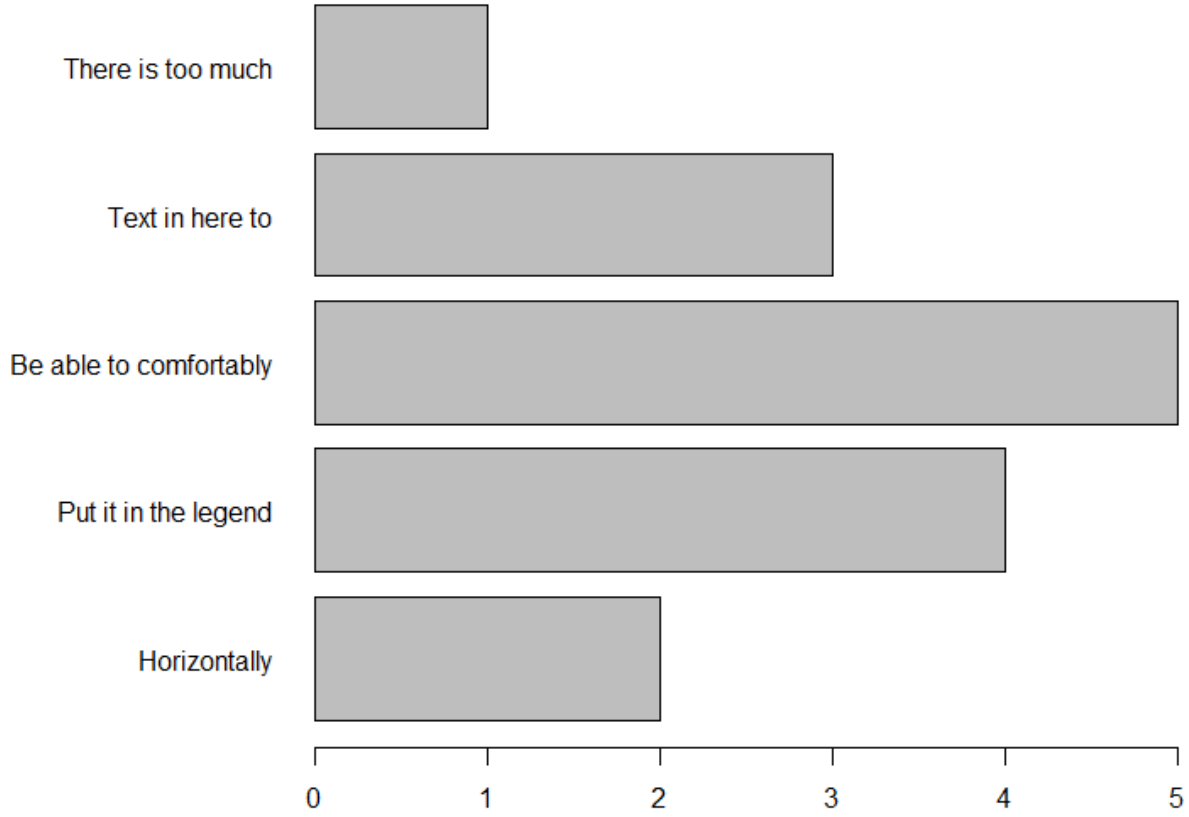
Keep text horizontal





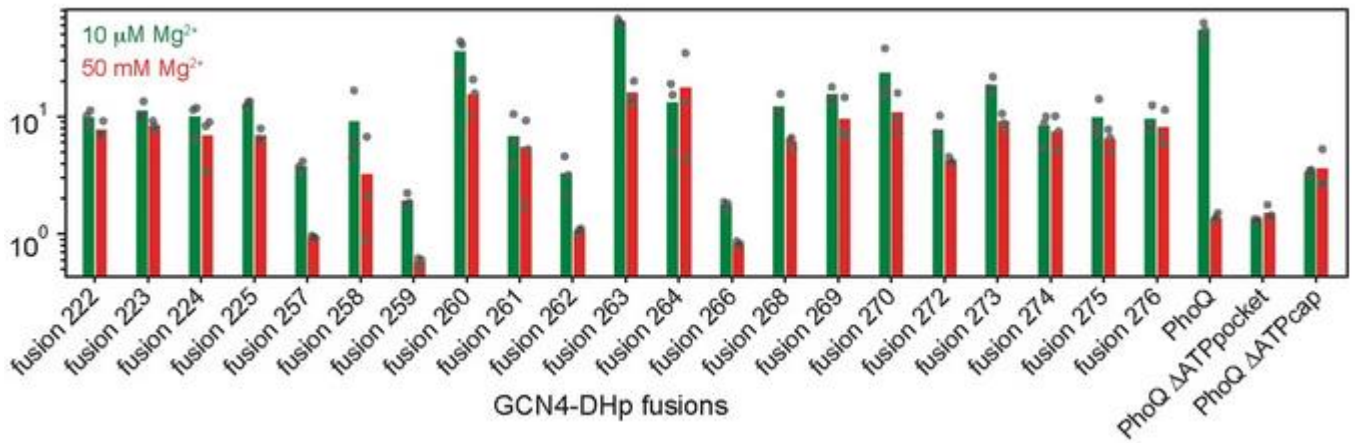
Keep text horizontal

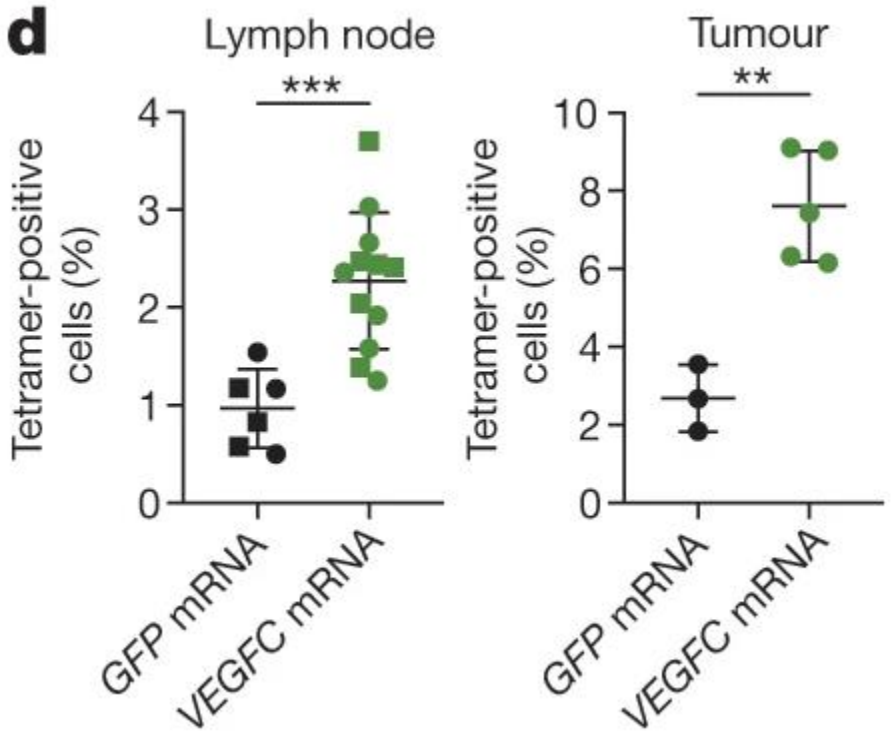




- Numbers are small, text is big
- All graphs still work when rotated 90^{o}

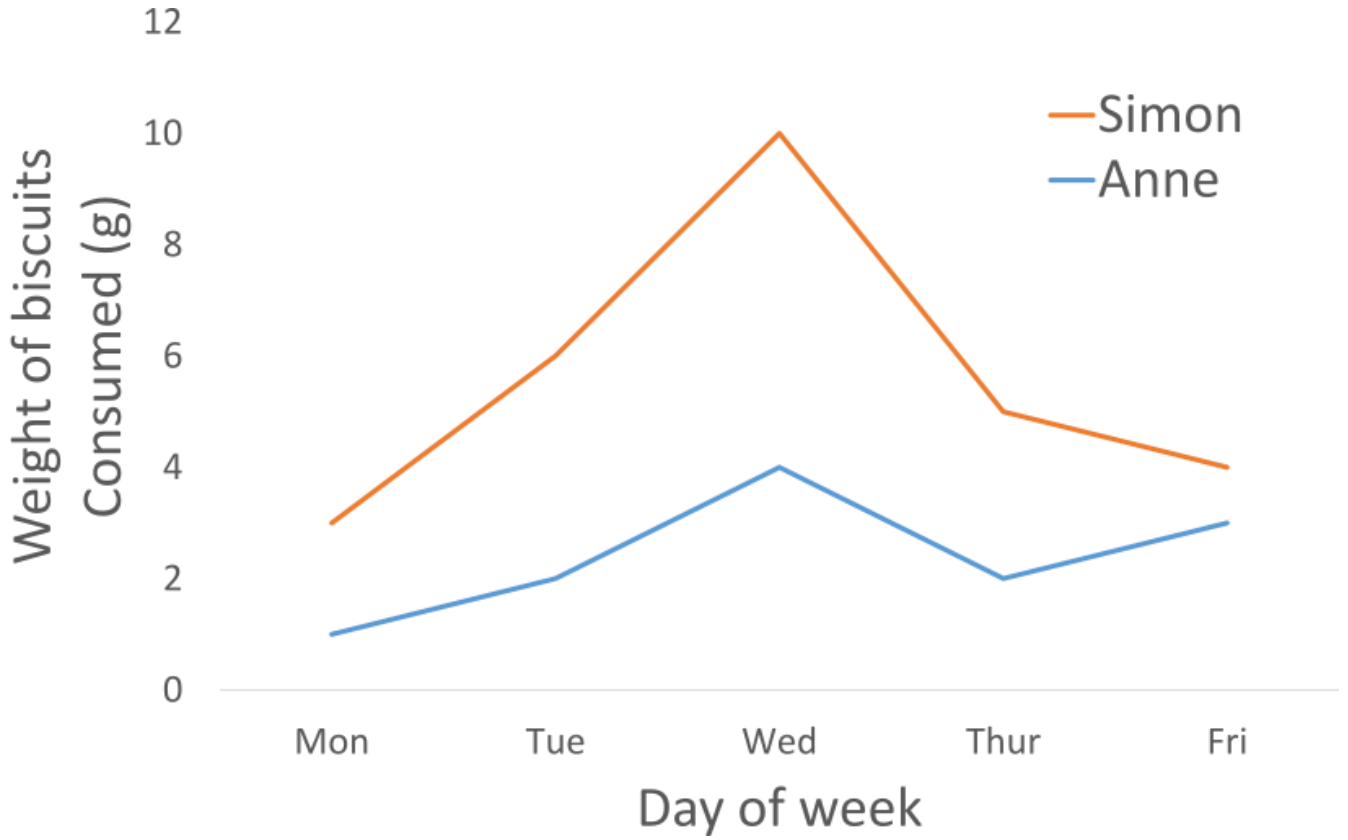
Keep text horizontal





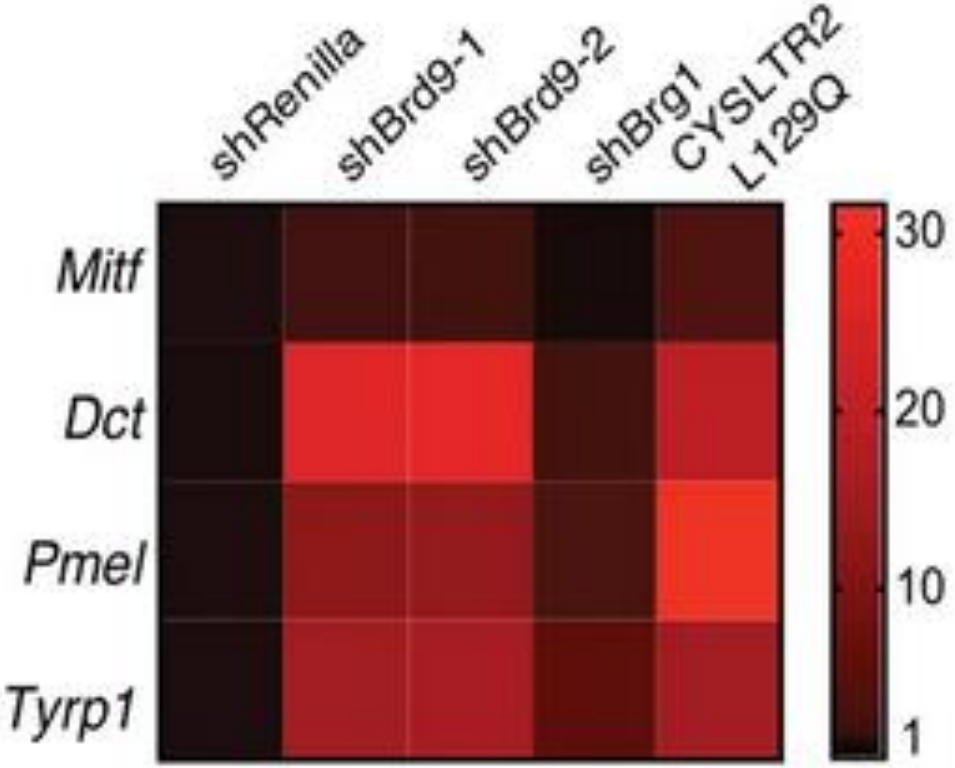
Labelling and annotation

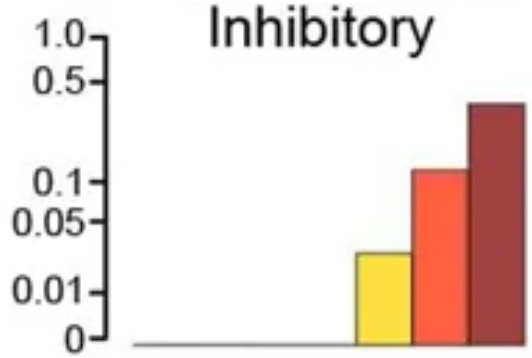
- Each axis is labelled
- Axis scales are appropriate
- Quantitative axes have units
- Colour scheme is explained
- Point shapes are explained

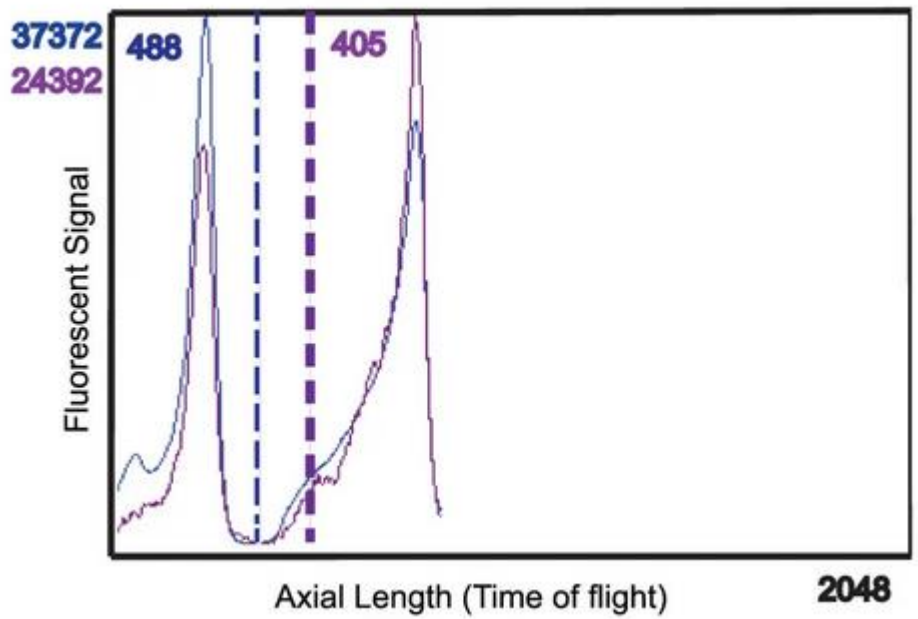


You need enough annotation that the figure is understandable on its own.

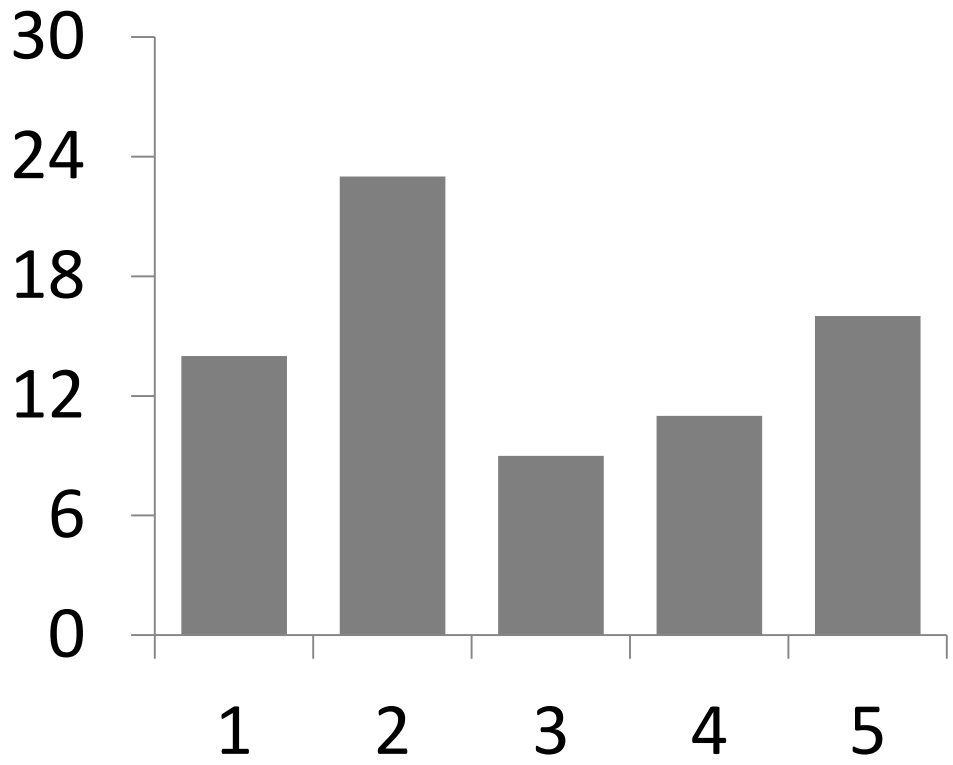
Labelling and annotation

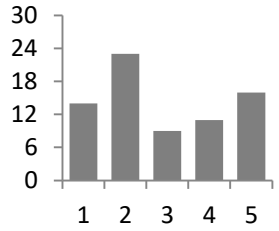






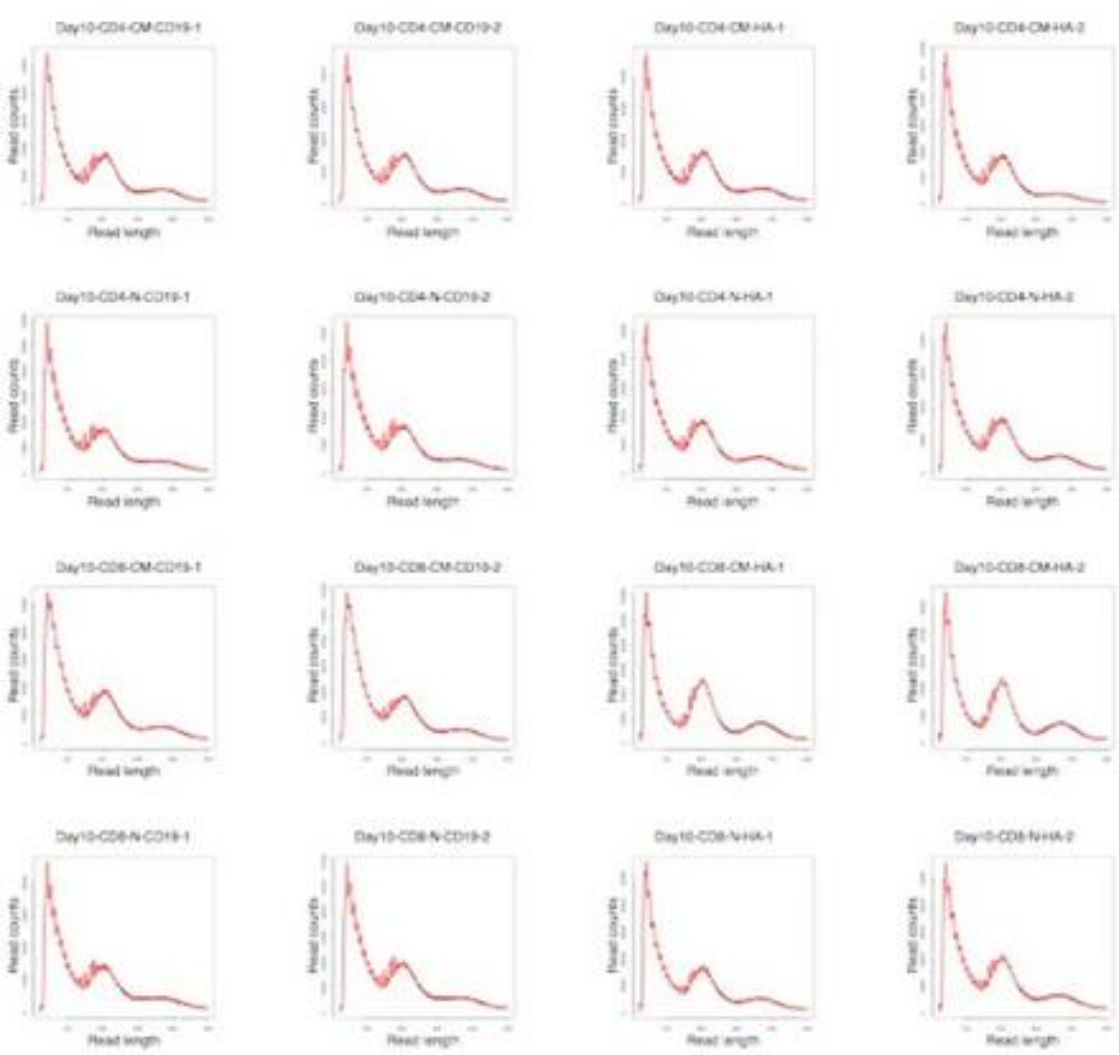
Make sure all text is legible
at the final printed size

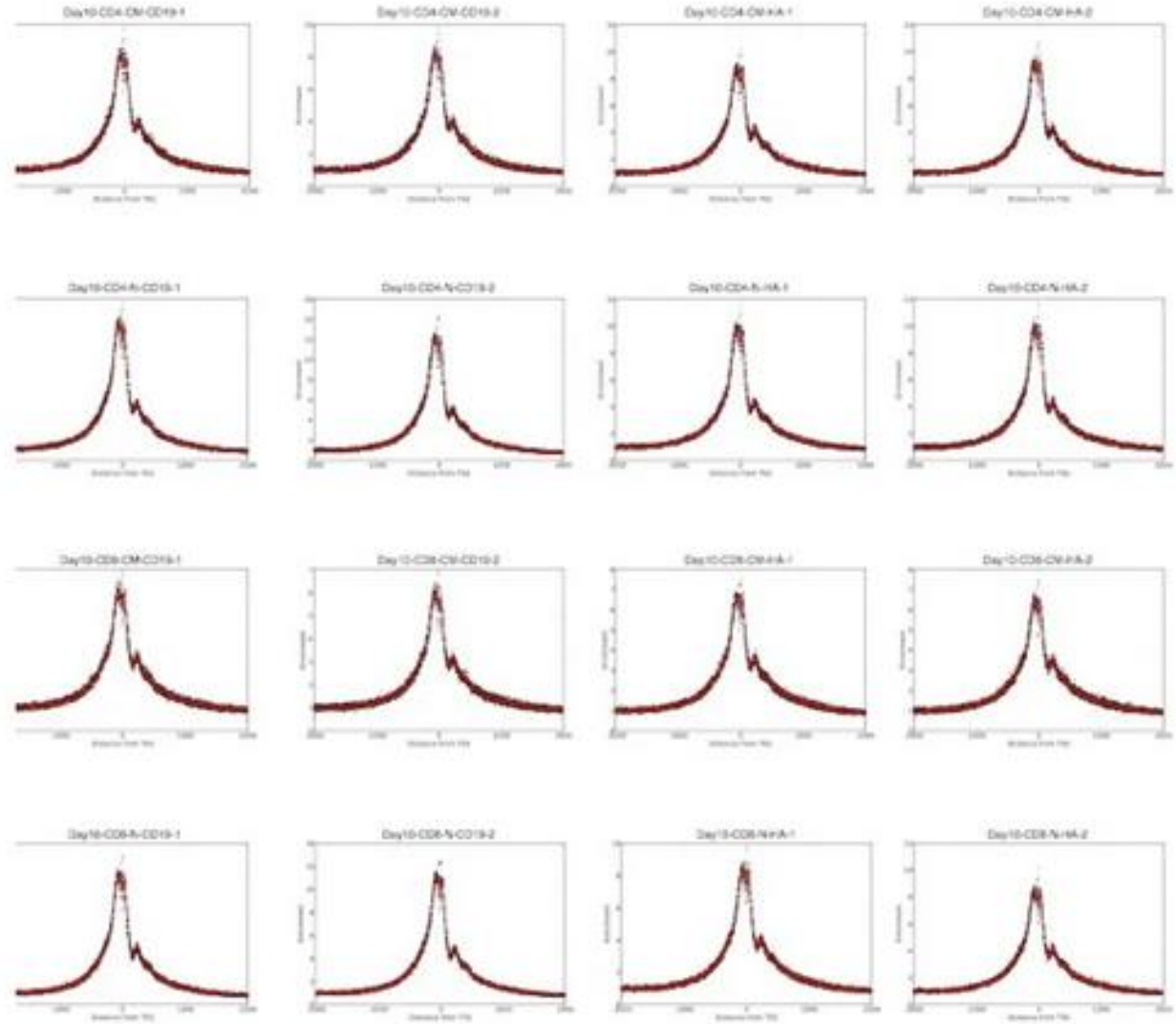




6 point font is the smallest you can comfortably read
(just over 2mm height on paper)

Make sure text is legible





When resizing be aware of what can and cannot have its aspect ratio changed

- Things that always need to maintain their aspect ratios:
  - Images
  - Text
  - Circular objects
  - Axes with comparable units

Checklist

- Consistent use of
  - Figure types
  - Colours / Shapes
  - Fonts and Sizes
  - Names
- Colour
  - Uses a standard scheme
  - Colourblind friendly (if possible)
- All figures are correctly annotated
  - Axes labelled with names and units
  - Colours and Shapes explained
- Text
  - Sans serif font
  - Large enough to be legible
  - Ideally in black or white
  - Sufficient contrast to be legible