Emerging Approaches for Tumor Analyses
in Epidemiological Studies

Session 12: Data Visualization

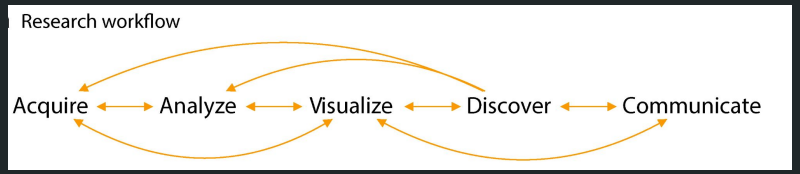
May 15, 2023

9:30 AM- 12:00 PM

Overview

- The Importance of Data Visualization in Scientific Research
- The Basics of Data Visualization
- Common Mistakes in Data Visualization
- Understanding Cancer Genetic Data and Creating Effective Visualizations
- Tools and Databases related to Data Visualization
- Suggestions for Creating Publication-Ready Figures
- Q&A and Introduction of Practical Session

The Importance of Data Visualization in Scientific Research



Why do we visualize data?

‘A Picture Is Worth a Thousand Words’

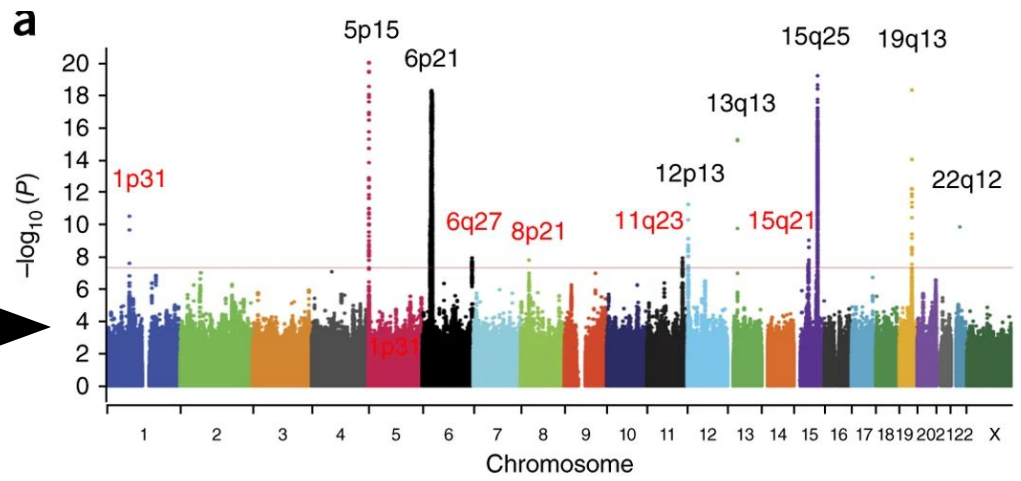
- Visualization can be helpful (or even essential) to represent results of statistical analyses, to formulate hypotheses and summarize theory, to explore your data so that you understand it better via exploratory analysis or outlier detection, and more.
- Scientific visuals can be essential for analyzing data, communicating experimental results and even for making surprising discoveries.
- Visualizations can reveal patterns, trends and connections in data that are difficult or impossible to find any other way
- Deficient data visuals can reduce the quality and impede the progress of scientific research.

Seeing is Believing: the Power of Data Visualization

Seeing is Believing: the Power of Data Visualization

Communicating complex data

Chr	Pos	P-value	….
1	12400	0.01	
1	12405	0.2	
1	12412	0.8	
….	….	….	

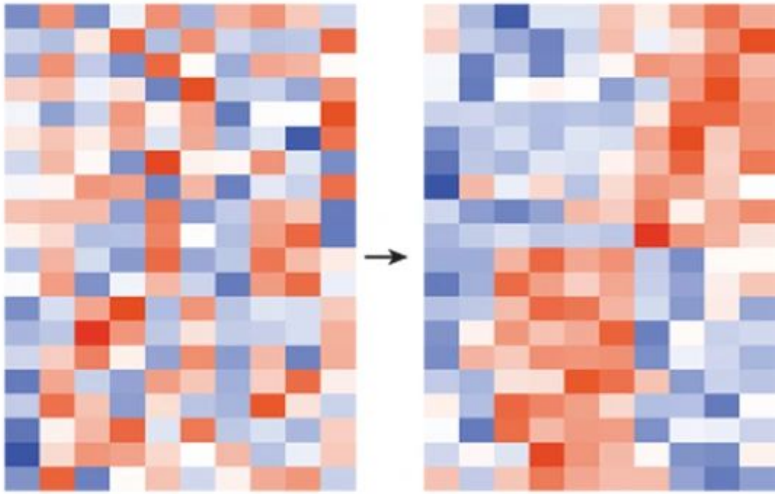


Manhattan plot

Seeing is Believing: the Power of Data Visualization

Enhancing data comprehension

Unclustered

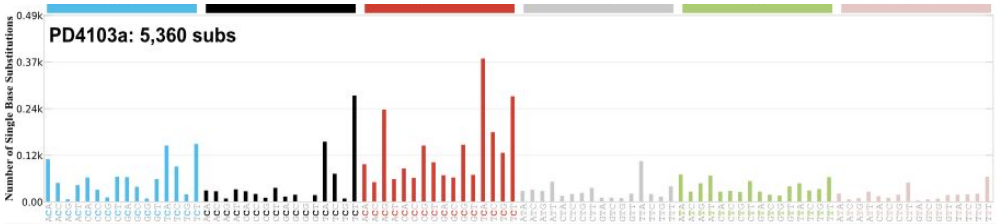


Clustered

Gene expression heatmap

Seeing is Believing: the Power of Data Visualization

Improving reproducibility

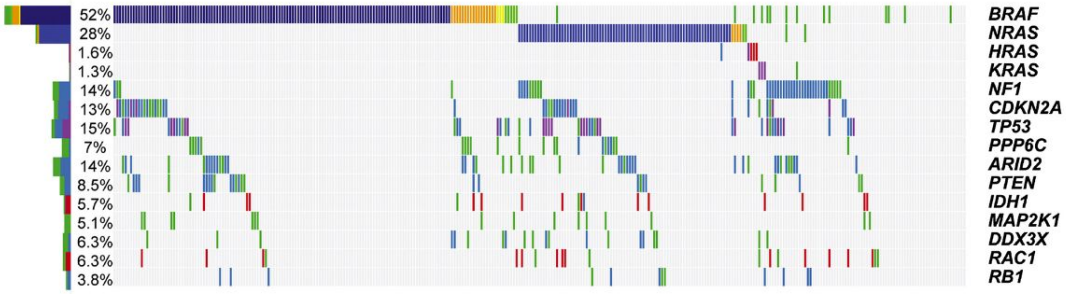


Mutational spectrum from one BRCA sample

Seeing is Believing: the Power of Data Visualization

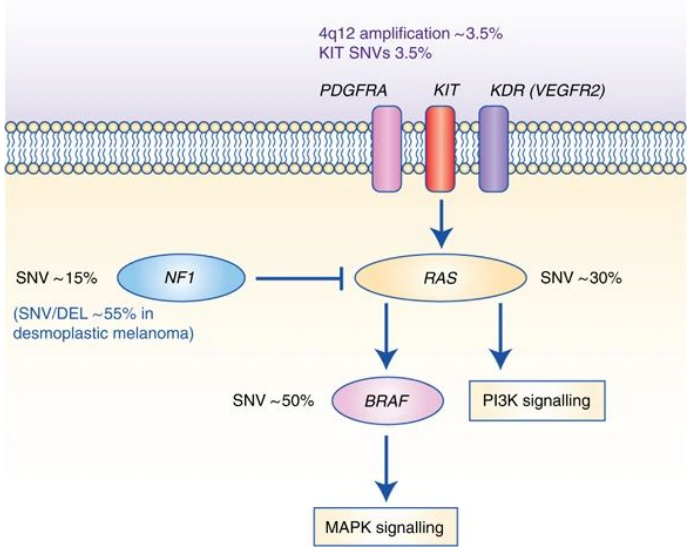
Facilitating interdisciplinary collaboration

Mutational exclusivity analysis from bioinformatician



Landscape of driver mutations in melanoma

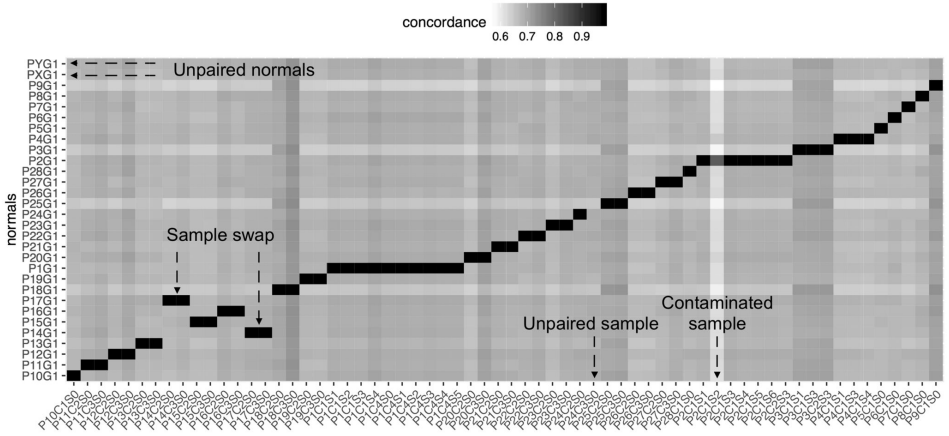
Proposed gene pathway from biologist



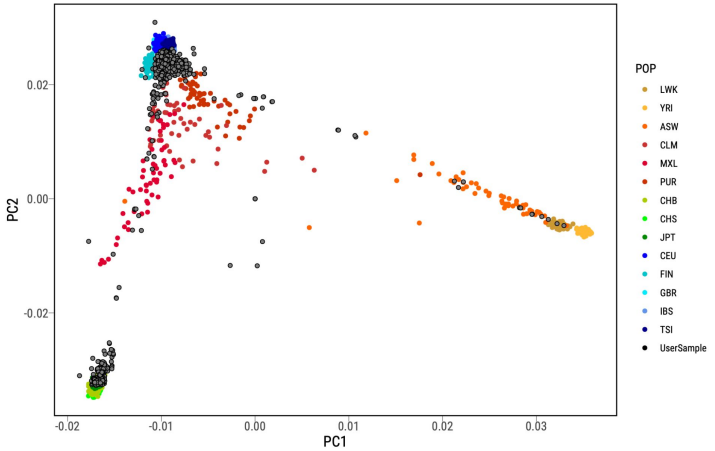
MAPK pathway genetic alterations in melanoma

Seeing is Believing: the Power of Data Visualization

Improving decision-making



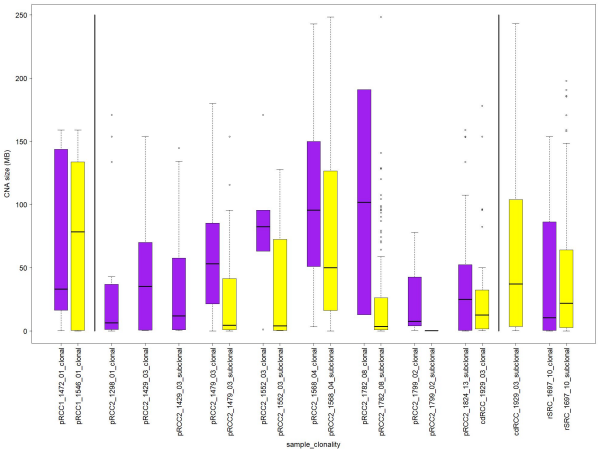
Rapid relatedness estimation for cancer and germline studies

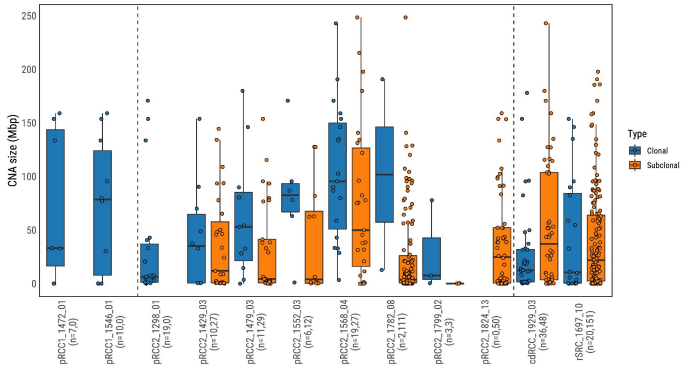


PCA analysis between user datasets and 1000 genome datasets

Increasing the likelihood of manuscript being accepted

A real story about a reviewer's comment related to data visualization





Comment from reviewer: ***The figures are stylish and nice-looking overall, but I found the purple/yellow combination in Figure 3b very brutal!***

Basics of Data Visualization

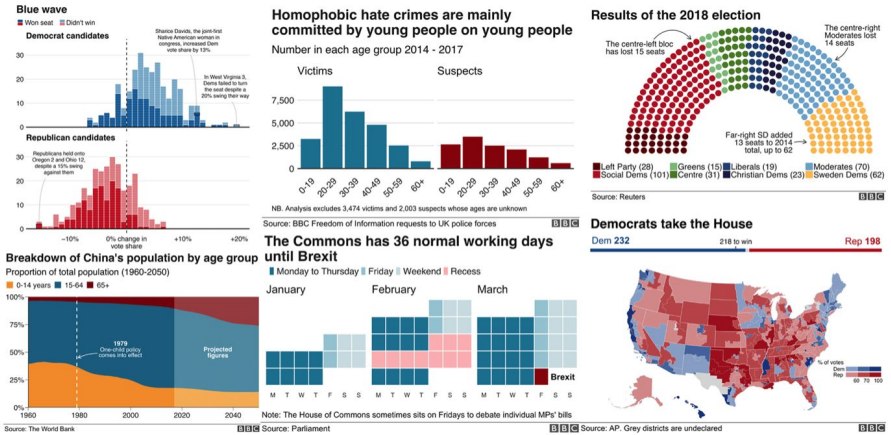
Examples based on R package ggplot2

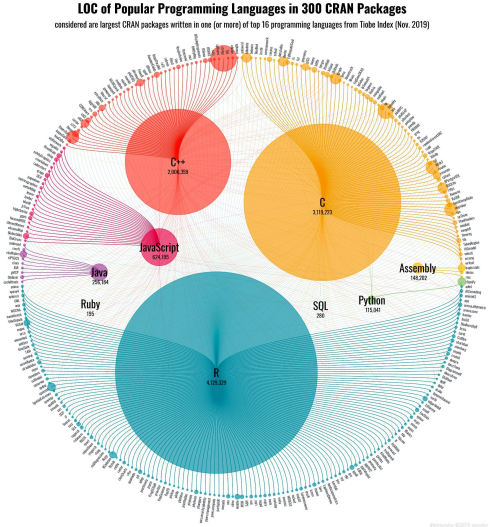
Platforms for Data Visualization

Programming-based	Web-based	Programming-based & Interactive	No programming required
ggplot2 matplotlib seaborn gnuplot	tableau Data-Driven Documents Datawrapper bio RENDER	Shiny plotly	Prism

How the BBC Visual and Data Journalism team works with graphics in R

BBC Visual and Data Journalism Follow
Feb 1 · 8 min read





CRAN R Packages by Number of Downloads

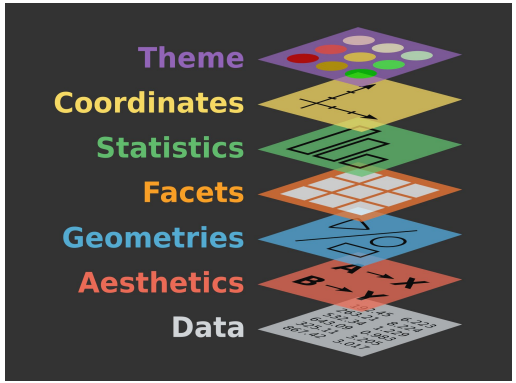
UPDATED DAILY. Last updated: 2023-05-03 00:14:42 +1000 Melbourne time (about 1 hour ago)

Rank	Package Name	Downloads	Author	Maintainer
1	ggplot2	111,438,231		
2	magrittr	109,570,938		
3	rlang	105,791,469		
4	dplyr	87,043,582		
5	vctrs	73,898,616		
6	tibble	73,058,604		
7	jsonlite	71,946,431		
8	cli	70,565,355		
9	Rcpp	70,539,069		
10	pillar	68,191,151		

Data Visualization R package: **ggplot2**

What is ggplot2?

The R package **ggplot2** is based on the Grammar of Graphics (GG), which is a framework for data visualization that dissects each component of a graph into individual components, creating distinct layers. Using the GG system, we can build graphs step-by-step for flexible, customizable results.




```
ggplot(data = <DATA>) +
  <GEOM function>(mapping=aes(<mappings>),
                  stat = <STAT>, position = <POSITION> ) +
  <COORDINATE function> +
  <SCALE function> +
  <THEME function> +
  <FACET function> +...
```

Required

Not required

To make a ggplot, the data and mapping layers are basic **requirements**, while the other layers are for additional customization. *The layers that are "not required" are still important to think about, but you will be able to generate a basic plot without them.*

Introduction to data visualization with ggplot2
https://r.qcbs.ca/workshop03/book-en/

- **Data**:
  - your data, in tidy format, will provide ingredients for your plot
  - use `dplyr` techniques to prepare data for optimal plotting format
  - usually, this means you should have one row for every observation that you want to plot
- **Aesthetics** (aes), to make data visible
  - `x`, `y`: variable along the x and y axis
  - `colour`: color of geoms according to data
  - `fill`: the inside color of the geom
  - `group`: what group a geom belongs to
  - `shape`: the figure used to plot a point
  - `linetype`: the type of line used (solid, dashed, etc)
  - `size`: size scaling for an extra dimension
  - `alpha`: the transparency of the geom
- **Geometric objects** (geoms - determines the type of plot)
  - `geom_point()`: scatterplot
  - `geom_line()`: lines connecting points by increasing value of x
  - `geom_path()`: lines connecting points in sequence of appearance
  - `geom_boxplot()`: box and whiskers plot for categorical variables
  - `geom_bar()`: bar charts for categorical x axis
  - `geom_histogram()`: histogram for continuous x axis
  - `geom_violin()`: distribution kernel of data dispersion
  - `geom_smooth()`: function line based on data
- **Facets**
  - `facet_wrap()` or `facet_grid()` for small multiples
- **Statistics**
  - similar to geoms, but computed
  - show means, counts, and other statistical summaries of data
- **Coordinates** - fitting data onto a page
  - `coord_cartesian` to set limits
  - `coord_polar` for circular plots
  - `coord_map` for different map projections
- **Themes**
  - overall visual defaults
  - fonts, colors, shapes, outlines

Detailed information can be found in the following website:

https://ggplot2.tidyverse.org/reference/index.html

ggplot2 theme elements

ggplot2
theme elements
reference

Set `minimal` as the baseline theme:

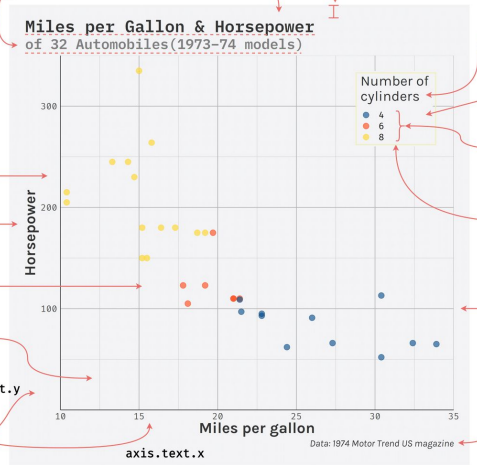
```
theme_minimal() +
theme(theme.element = element_type())
```

Use `element_blank()` to **remove an element**

Axis titles, **text**, **ticks**, and **lines** can be specified per axis using theme inheritance by putting `.x`/`.y` at the end of the theme element.

`plot.title.position = "plot"`
`plot.caption.position = "plot"`
"plot" means that they will be aligned to the entire plot (instead of the panel)

`plot.title = element_text()`
`plot.subtitle = element_text()`
`plot.margin = margin(25, 25, 25, 25)`



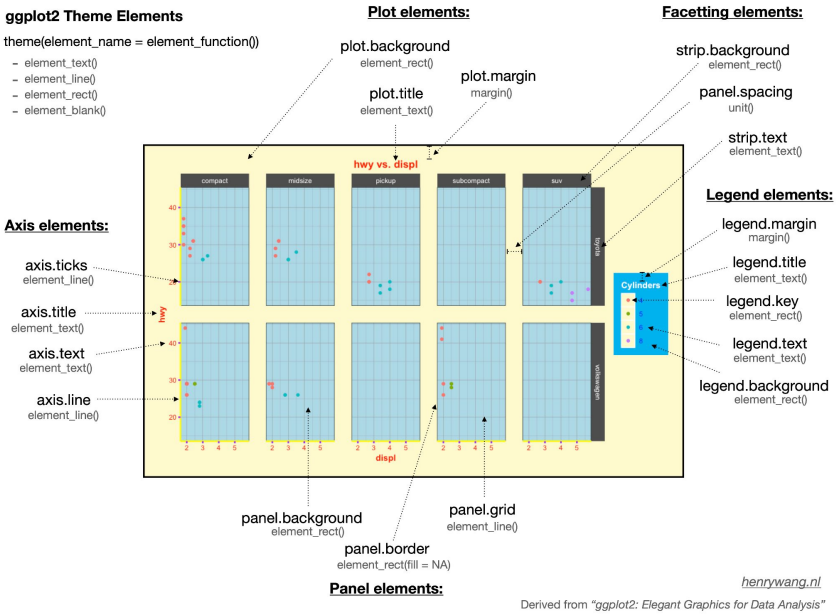
`legend.title = element_text()`
`legend.background = element_rect()`
`legend.text = element_text()`
`legend.position = c(.85,.85) / "none" / "left" / "right" / "bottom" / "top"`
`plot.background = element_rect()`
`plot.caption = element_text()`

`axis.line.y = element_line()`
`axis.title.y = element_text()`
`panel.grid.major = element_line()`
`panel.grid.minor = element_line()`
`axis.text.y`
`axis.text = element_text()`
`axis.text.x`

`text = element_text()` ← modifications will be applied to all text elements

isabella-b

Full list of elements at ggplot2.tidyverse.org/reference/theme



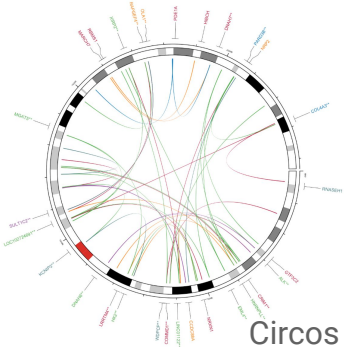
henrywang.nl

Derived from *"ggplot2: Elegant Graphics for Data Analysis"*

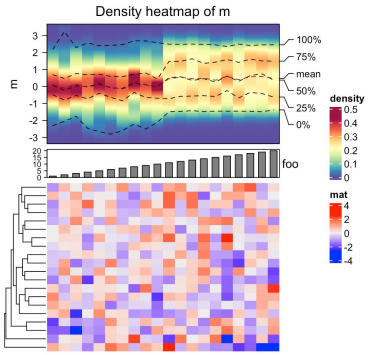
Plot Type

Simple plot

Complex plot

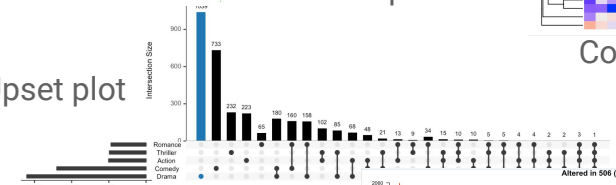


Circos plot

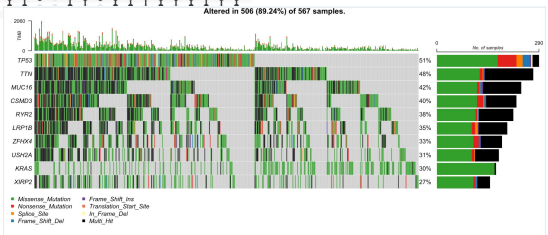


ComplexHeatmap

Upset plot

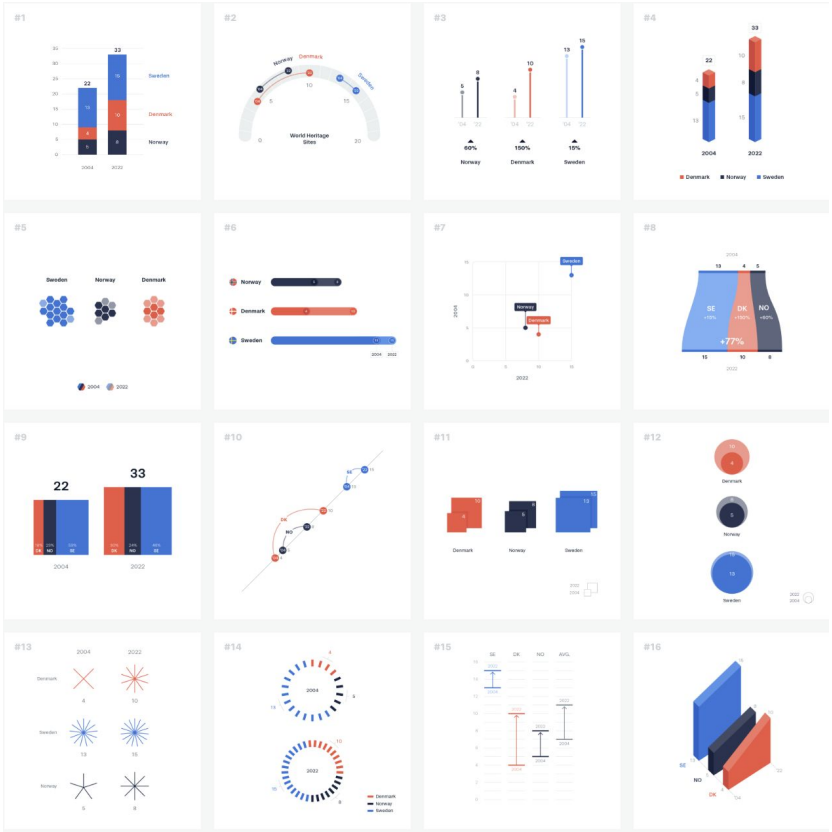


Oncoplot



Inspired by “1 dataset 100 visualizations”

https://100.datavizproject.com/



Box plots are a more communicative way to show sample data

nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › correspondence › article

Published: 30 January 2014

BoxPlotR: a web tool for generation of box plots

Michaela Spitzer, Jan Wildenhain, Juri Rappsilber ✉ & Mike Tyers ✉

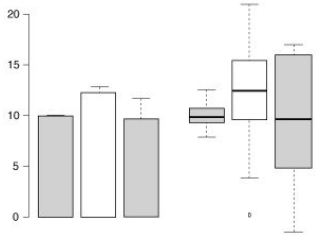
Nature Methods **11**, 121–122 (2014) | Cite this article

58k Accesses | **461** Citations | **127** Altmetric | Metrics

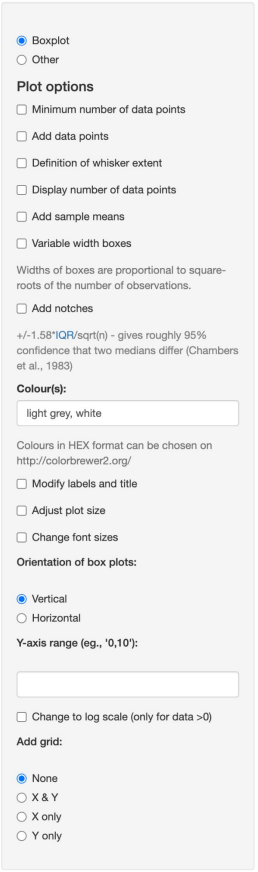
To the Editor

In biomedical research, it is often necessary to compare multiple data sets with different distributions. The bar plot, or histogram, is typically used to compare data sets on the basis of simple statistical measures, usually the mean with s.d. or s.e.m. However, summary statistics alone may fail to convey underlying differences in the structure of the primary data (Fig. 1a), which may in turn lead to erroneous conclusions. The box plot, also known as the box-and-whisker plot, represents both the summary statistics and the distribution of the primary data. The box plot thus enables visualization of the minimum, lower quartile, median, upper quartile and maximum of any data set (Fig. 1b). The first documented description of a box plot–like graph by Spear[1] defined a range bar to show the median and interquartile range (IQR, or middle 50%) of a data set, with whiskers extended to minimum and maximum values. The most common implementation of the box plot, as defined by Tukey[2], has a box that represents the IQR, with whiskers that extend 1.5 times the IQR from the box edges; it also allows for identification of outliers in the data set. Whiskers can also be defined to span the 95% central range of the data[3]. Other variations, including bean plots[4] and violin plots, reveal additional details of the data distribution. These latter variants are less statistically informative but allow better visualization of the data distribution, such as bimodality (Fig. 1b), that may be hidden in a standard box plot.

https://www.nature.com/articles/nmeth.2811

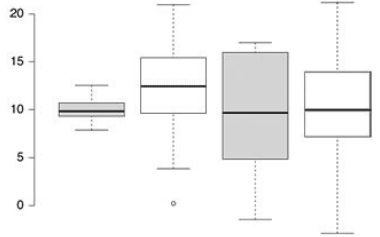


The same three samples plotted by bar chart with s.e.m. error bars (**left**) and Tukey-style box plot (**right**). The box plot more clearly represents the underlying data.



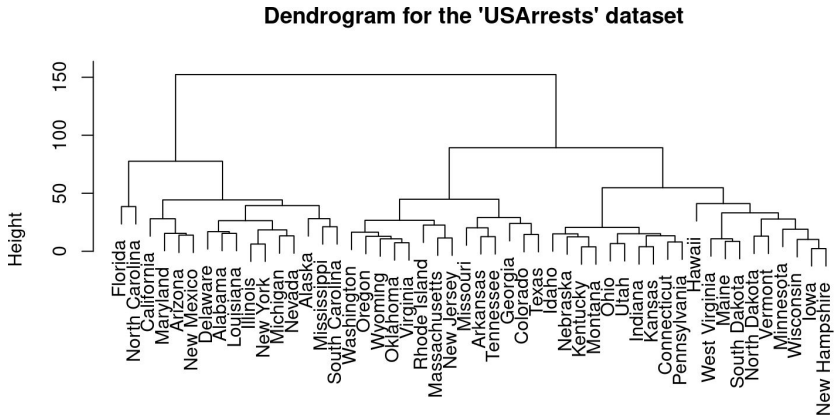
boxplotR

http://shiny.chemgrid.org/boxplotr/

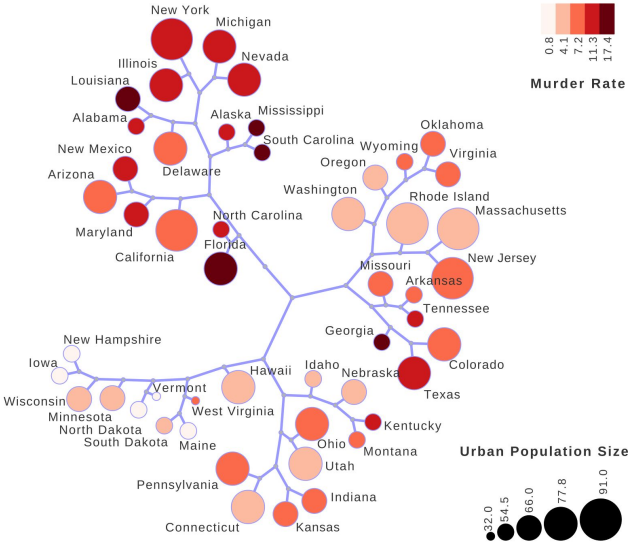


Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. n = 100, 76, 16, 76, 41 sample points.

Choose the best visualization for your data



Dendrogram diagram displays binary trees focused on representing hierarchical relation

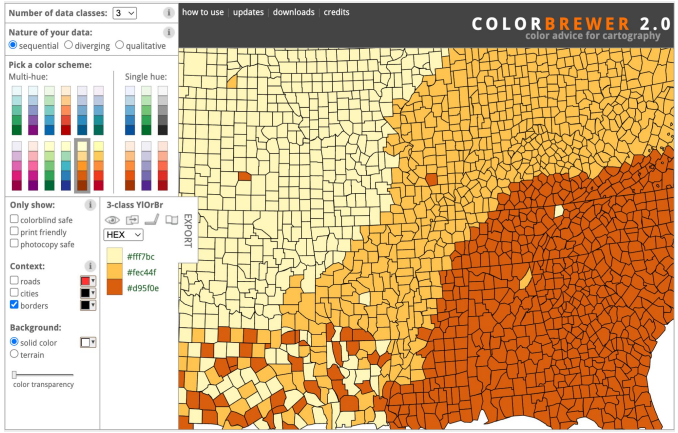


TreeAndLeaf Plot

http://bioconductor.org/packages/release/bioc/vignettes/TreeAndLeaf/inst/doc/TreeAndLeaf.html

Colors

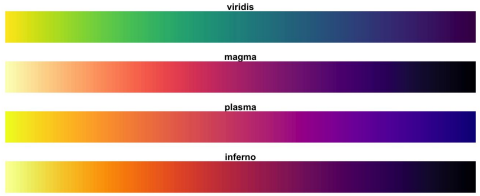
- Highlight important information
- Distinguish between data categories
- Create contrast for easier reading
- Convey emotion or mood
- Enhance aesthetic appeal
- Convey quantitative information
- **Use colors carefully to avoid confusion.**



https://colorbrewer2.org/

ggsci

NPG/AAAS/NEJM/Lancet/JAMA/JCO/Frontiers
LocusZoom/IGV/COSMIC/GSEA
……..



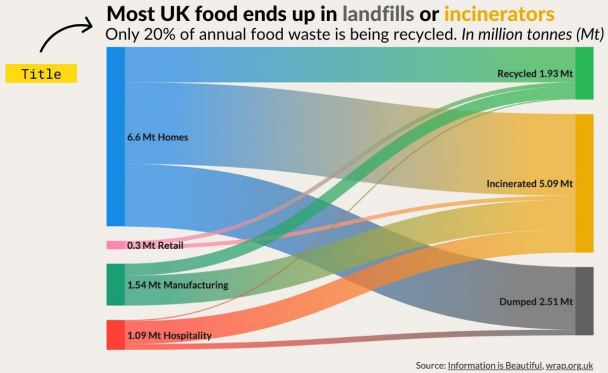
Viridis palette
(Colorblind-Friendly)

#BB0E3D	#ff7f00	#4daf4a	#007BBD	#984ea3
#467E84	#14315C	#3D4551	#947100	#4BBFC6
#FACE00	#a65628	#f781bf	#71767A	#b2df8a
#cab2d6	#ffff33	#542788	#bababa	#01665e

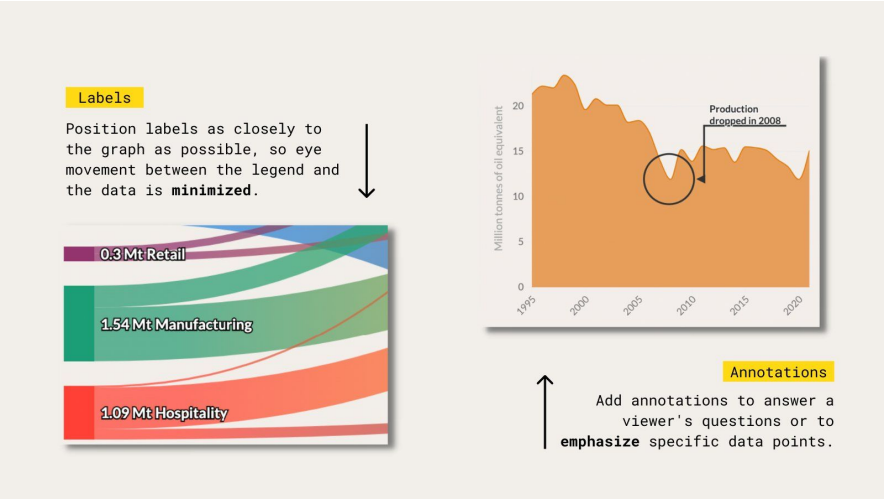
NCI Color Palette

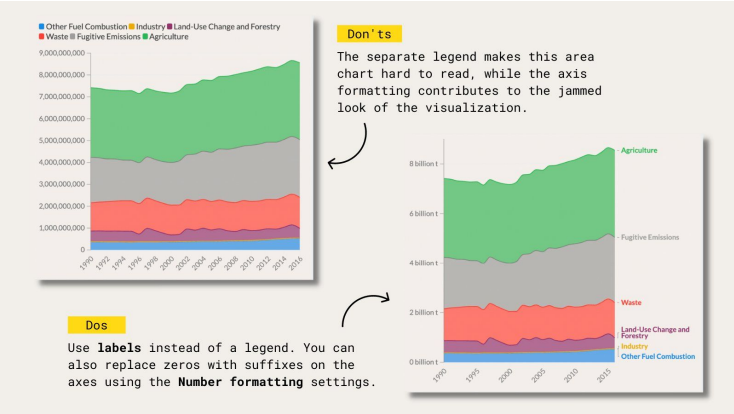
Annotations/Text

Text is one of the most crucial elements of data visualization. Here are a few useful tips about using text in your data visualization:



Since graphs do not contain much text, existing text must be concise and effective. A **6-12 word** descriptive title is enough. You can also use **colors in the header** – read more in our help docs.



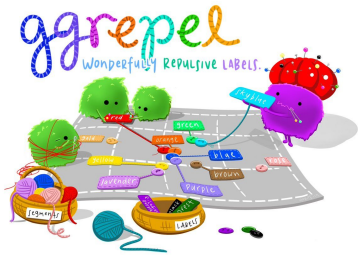


Fonts

- Some fonts are very easy to read, SOME CAN HURT TO READ
- What makes a good font
  - Taller letters
  - Rounded o's, p's, d's, etc.
  - Simple, un-elaborate letter shape
- Good example fonts:
  - Roboto
  - Sans
  - Lora
  - Etc.
- More info:
  - Choosing Fonts for Your Data Visualization
  - Which fonts to use for your charts and tables

R packages: showtext/extrafont

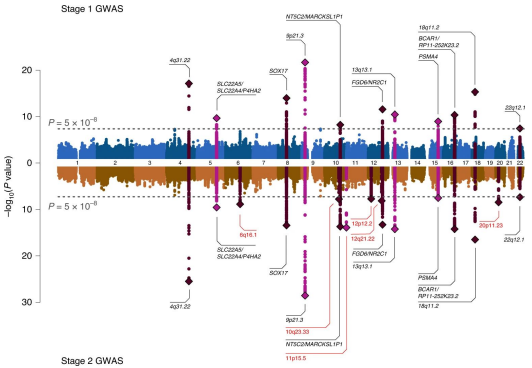
Labels

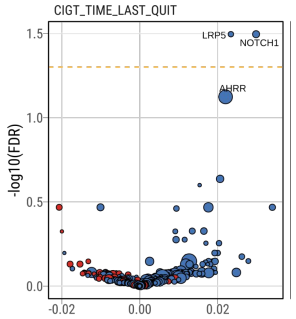


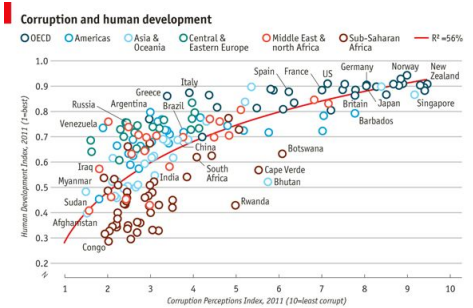
ggrepel provides geoms for ggplot2 to repel overlapping text labels:

- geom_text_repel()
- geom_label_repel()

Text labels repel away from each other, away from data points, and away from edges of the plotting area.



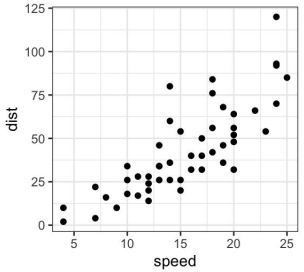




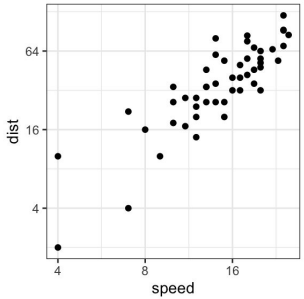
Scales/Transformation

You can construct your own transformer using scales::trans_new(), but ggplot2 understands many common transformations supplied by the scales package.

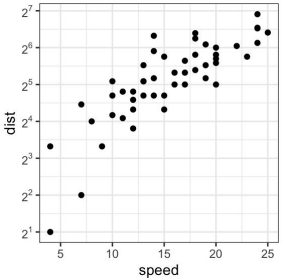
Name	Function $f(x)$	Inverse $f^{-1}(y)$
asn	$\tanh^{-1}(x)$	$\tanh(y)$
exp	e^x	$\log(y)$
identity	x	y
log	$\log(x)$	e^y
log10	$\log_{10}(x)$	10^y
log2	$\log_2(x)$	2^y
logit	$\log(\frac{x}{1-x})$	$\frac{1}{1+e(y)}$
pow10	10^x	$\log_{10}(y)$
probit	$\Phi(x)$	$\Phi^{-1}(y)$
reciprocal	x^{-1}	y^{-1}
reverse	$-x$	$-y$
sqrt	$x^{1/2}$	y^2



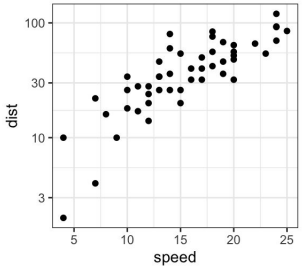
Log2 transformation of x and y axes



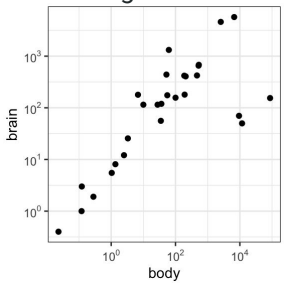
Format ticks label to show exponents



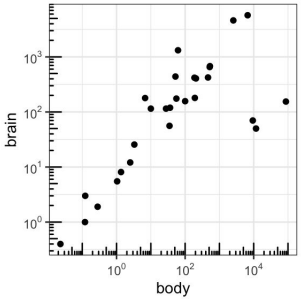
Set axis into log10 scale



Show log scale ticks



Display log scale ticks mark

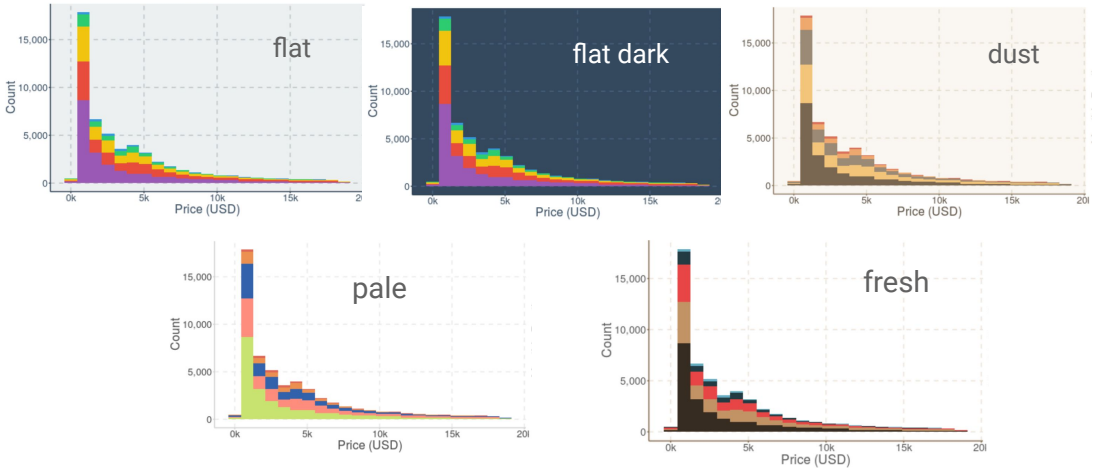


Theme

- Enhance visual appeal
- Unify visualizations
- Reinforce message/story
- Create visual hierarchy
- **Impact overall effectiveness.**

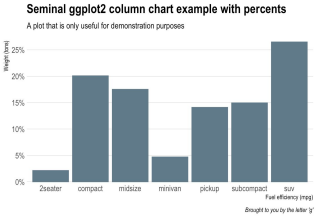
ggthemr

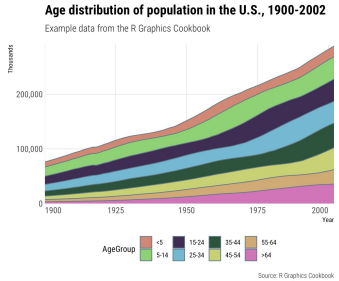
1) Colour palette; 2) Layout of axes lines and gridlines; 3) Spacing around plot and between elements

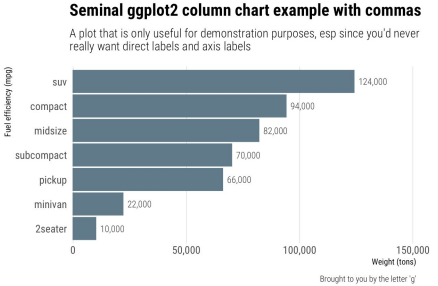


https://github.com/Mikata-Project/ggthemr

hrbrthemes provides typography-centric themes and theme components for ggplot2

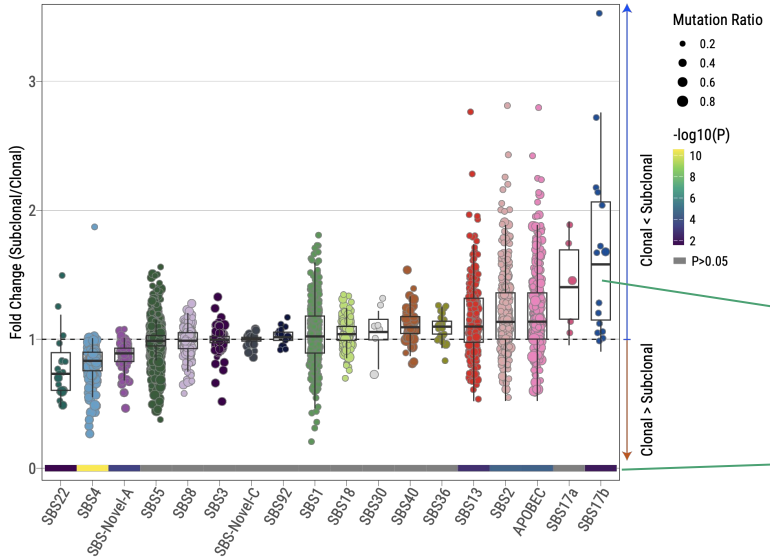


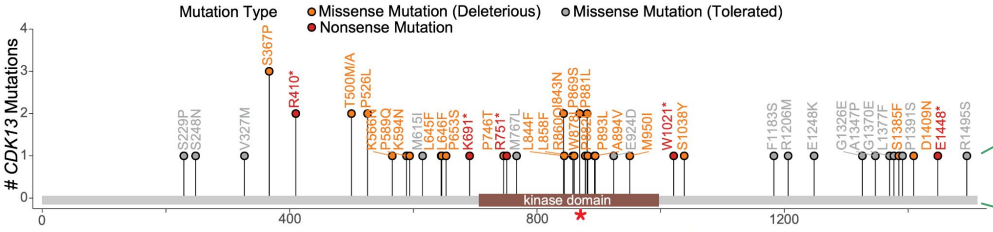




https://github.com/hrbrmstr/hrbrthemes

Data => Visualization





- **Multiple datasets can be visualized at the same time**
- **Additional information can be included as decorative 'data' in a visualization**

Data from boxplots come from the same data source with the same format

Different data for p-values

Points come from real data

Rectangles created for purpose

Additional resources

https://ggplot2-book.org/

https://r-charts.com/

https://python-charts.com/

https://cedricscherer.netlify.app/2019/08/05/a-ggplot2-tutorial-for-beautiful-plotting-in-r/

Fundamentals of data visualization- Wilke (free book): https://clauswilke.com/dataviz/directory-of-visualizations.html

Data Visualization- Healy (free book): https://socviz.co/

R for Data Science- Wickham (free book): https://r4ds.had.co.nz/

Common Mistakes in Data Visualization

Using the wrong type of visualization for the data

A good data visualization should be done:

- Clearly
- Precisely and accurately
- Effectively and efficiently
- excluding unnecessary elements: Including unnecessary or distracting elements, such as 3D effects or excessive ornamentation, can detract from the clarity of the visualization.

To ensure you are communicating effectively, here are some important questions to ask yourself when visualizing your data:

- What do you want to communicate?
- Who is your audience?
- What is the best way to visualize your message?

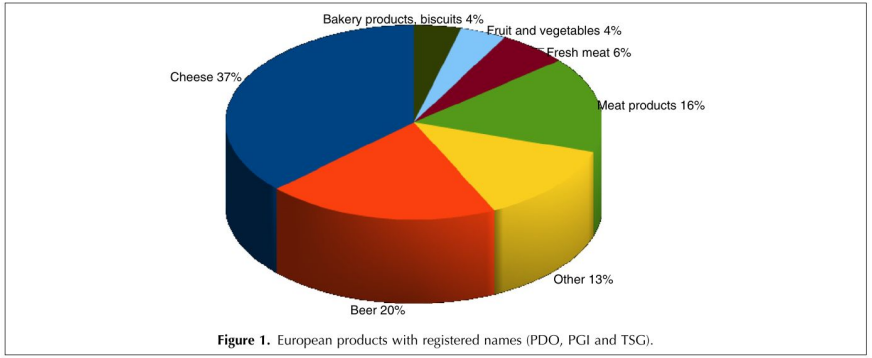


Figure 1. European products with registered names (PDO, PGI and TSG).

An example of an ineffective visualization

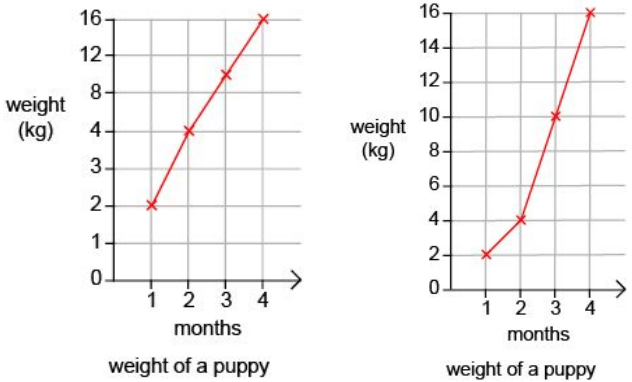
What is wrong with this image? Try to answer for yourself what you think this figure is trying to communicate and whether it is successful. How could this image be improved?

Hints: Why is this in **3D**? Which elements do your eyes **focus on** first? How does a **pie chart** visually communicate the information?

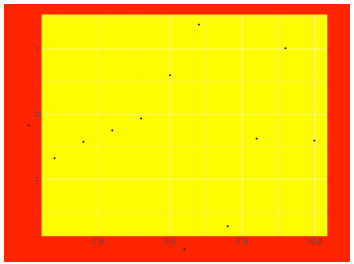
Using misleading scales or strong background

Tripled house price?

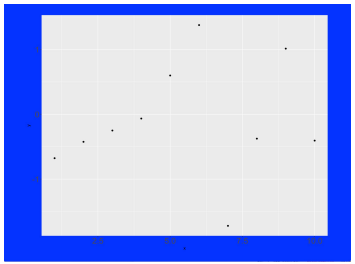
Greatest increase in weight between month 1 and 2?



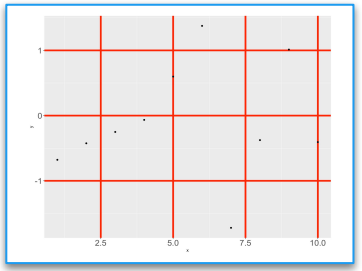
Avoid backgrounds that clash with the data



Avoid distracting backgrounds, which is hard for readers to focus on the data



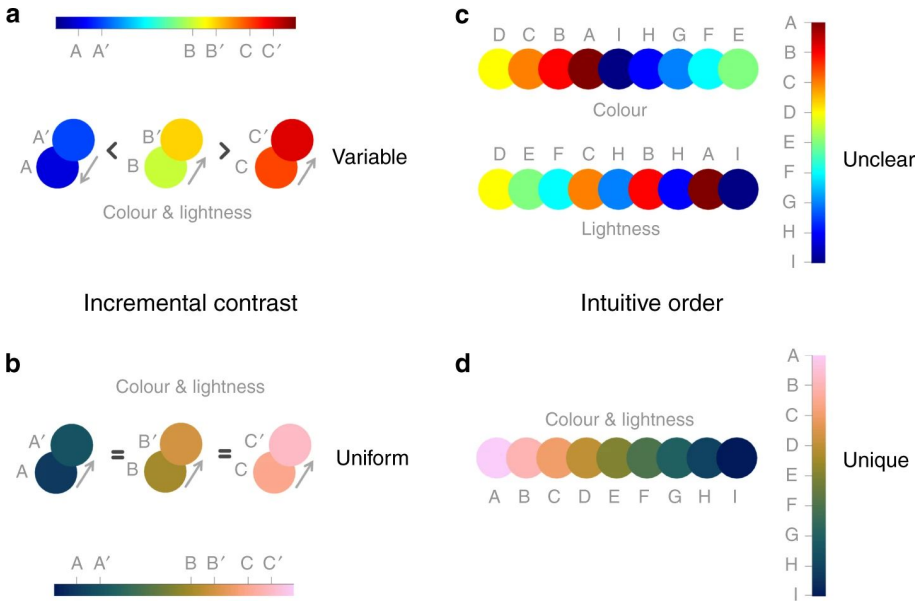
Avoid using gridlines that are too prominent

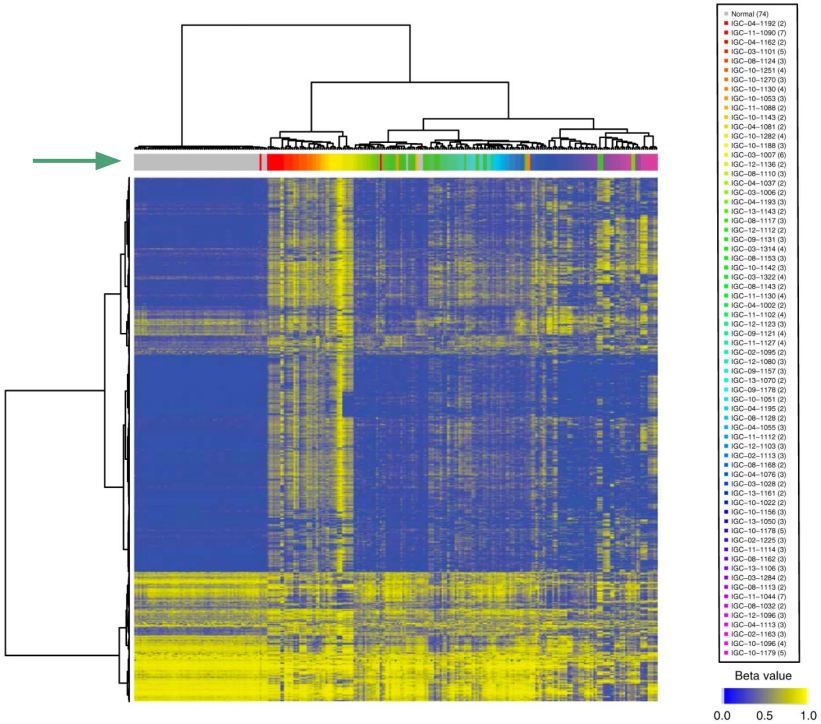


The misuse of colour in science communication

Poorly chosen color schemes or color scales can obscure the data or create confusion

Certain incremental data variation is either under- or strongly overrepresented with jet (a.k.a. rainbow) depending on the colour map segment



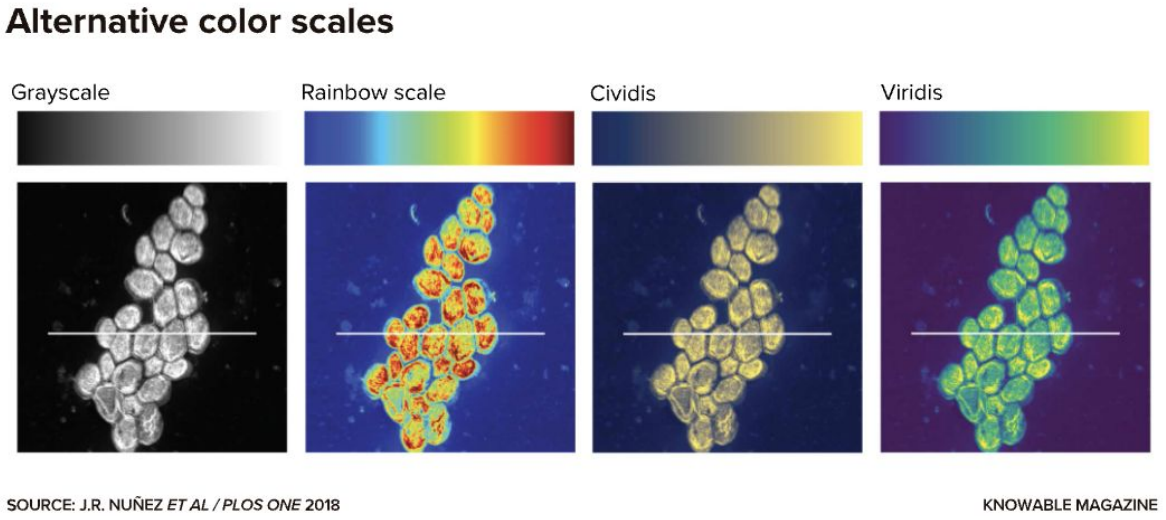


https://www.nature.com/articles/s41467-020-19160-7

Ruinous rainbows

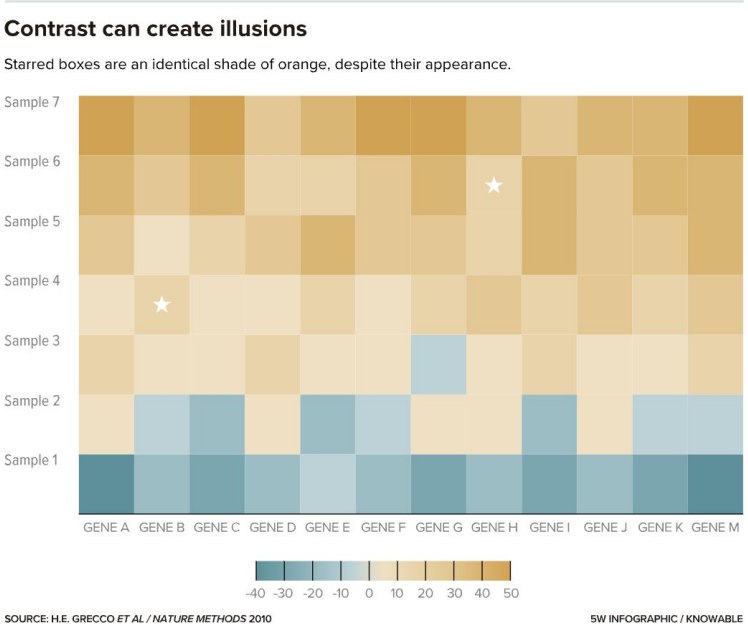
One of the most common bad practices is using the rainbow color scale. From geology to climatology to molecular biology, researchers gravitate toward mapping their data with the help of Roy G. Biv.

But the rainbow palette has several serious drawbacks — and very little to recommend it.

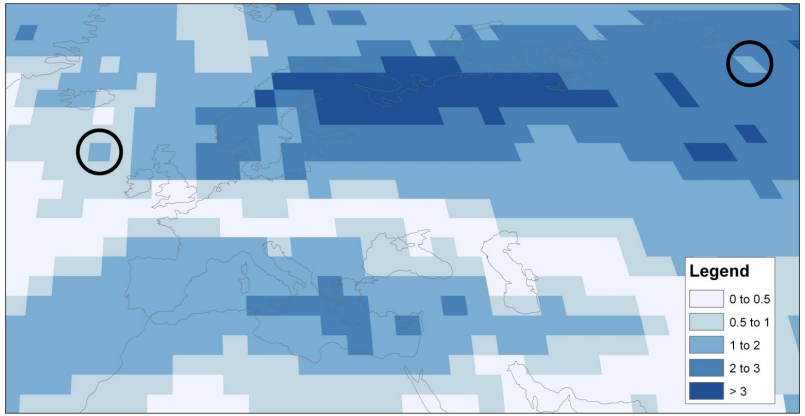


A microscopic image of yeast cells rendered with different color scales highlights the counterintuitive nature of the rainbow scale. Both the viridis and cividis color scales are intended to better represent the underlying data and are easier to read. Cividis was specifically designed to be legible for color-blind people as well.

Contrast can create illusions

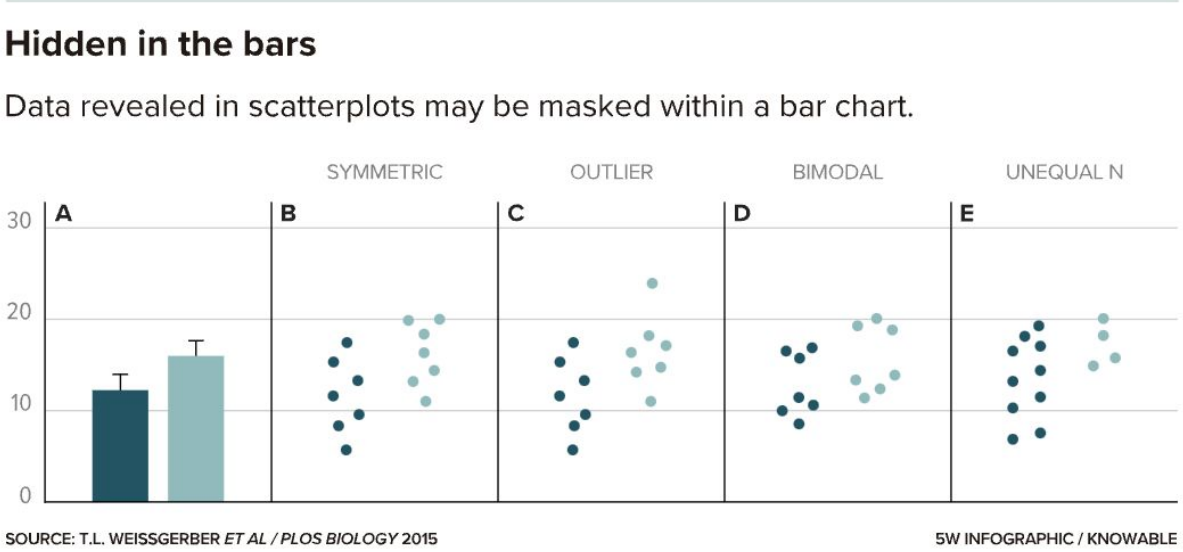


The two starred squares on this heat map are identical shades of orange, indicating identical values in terms of gene activity. But differences in the color of neighboring squares means that the starred ones don't look identical, which can be misleading.



Simultaneous contrast illusion means that the grid cell circled on the right hand side does not appear the same colour as the one on the left hand side. Also the grid cell on the right hand side does not match the legend (even though they are both in the same 1–2 category).

Data may be hidden in barplots

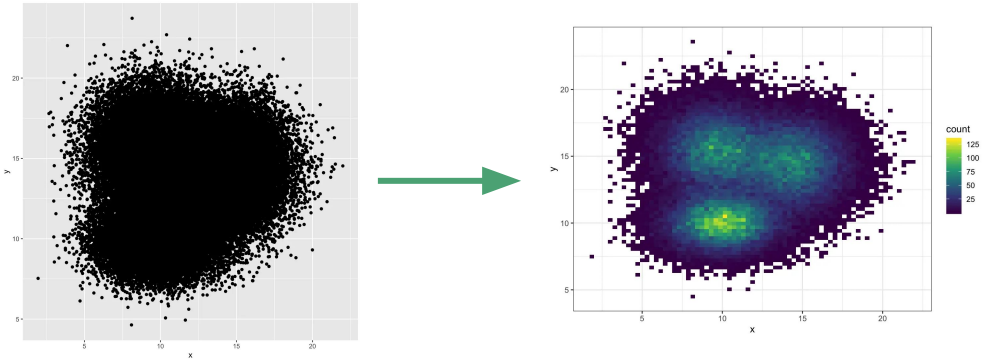


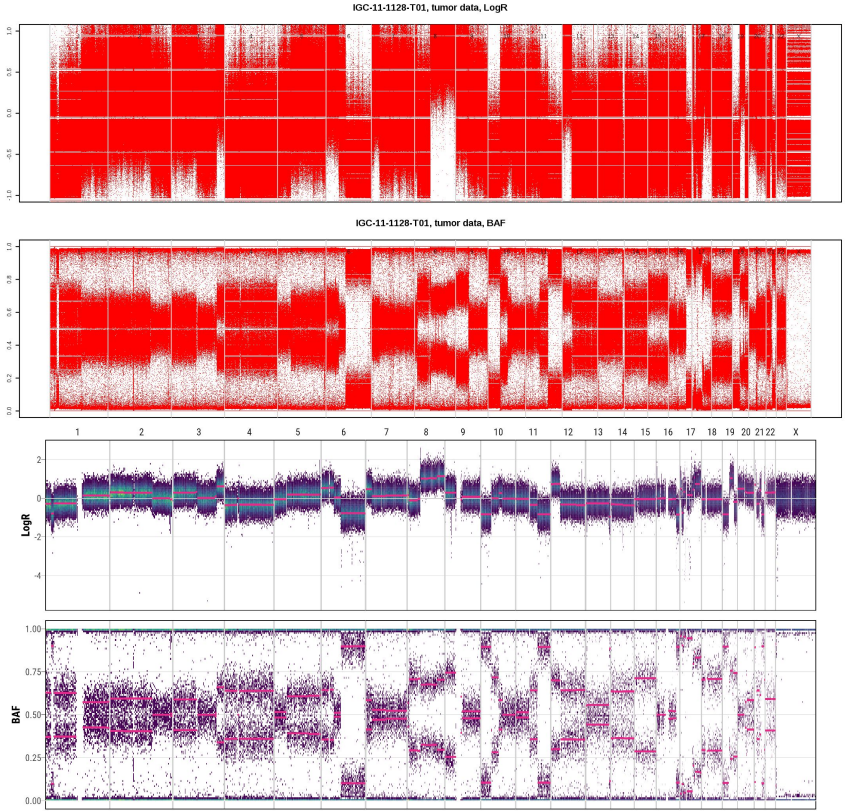
Box plots with data points are a more communicative way to replace the bars

Every one of the four sets of data on the right can be accurately represented by the same bar graph on the left, illustrating how bar graphs can obscure important details about the data, possibly misleading readers.

Cluttering the visualization

Including too much information or too many data points in a visualization can make it difficult to read and interpret.

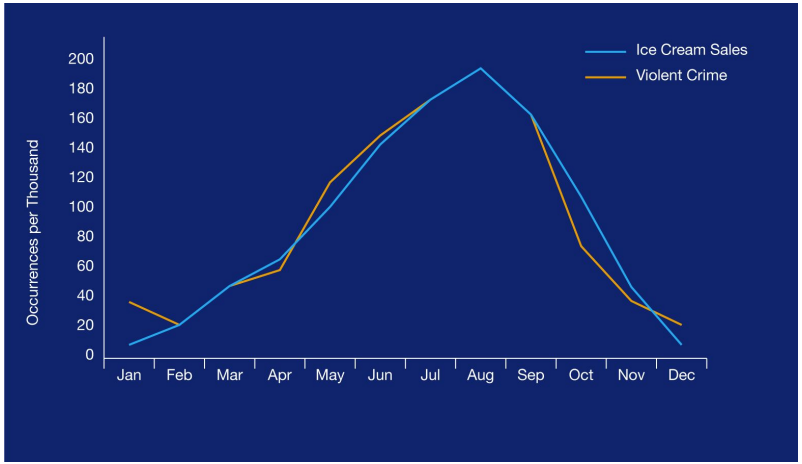




Copy number segmentation (logR + BAF)

Confusing Correlations

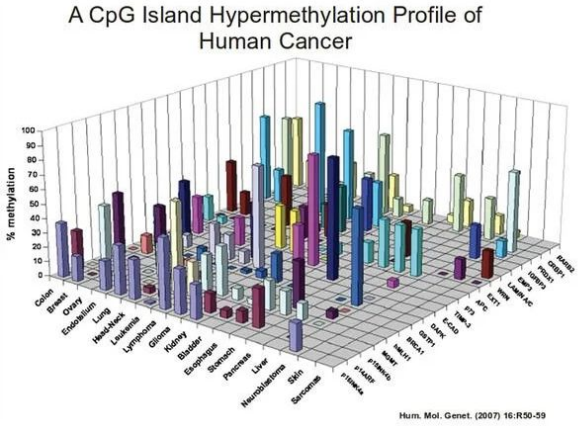
It's also possible to visualize correlations in a way that falsely implies causation. A famous example is linking increased ice cream sales to surges in violent crime when both are results of warm weather.

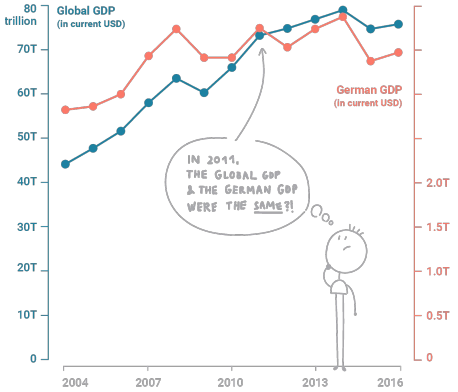


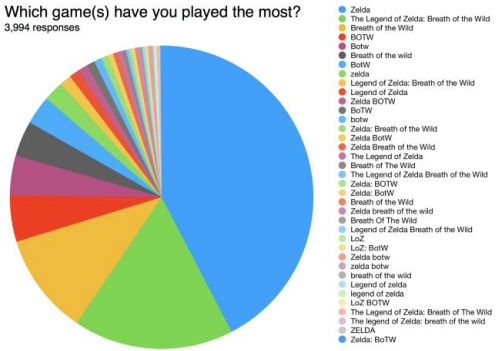
- It can be helpful to highlight correlations with multiple visualizations that exist in close proximity. This allows viewers to assess the data and still make connective links.
- It's worth restating. **Correlation doesn't equal causation**.

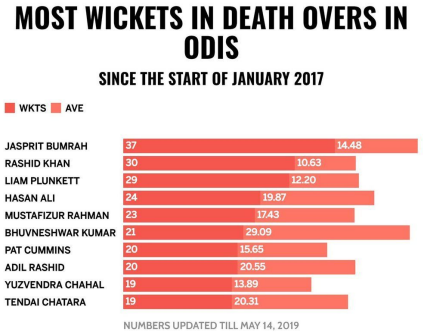
Other mistakes

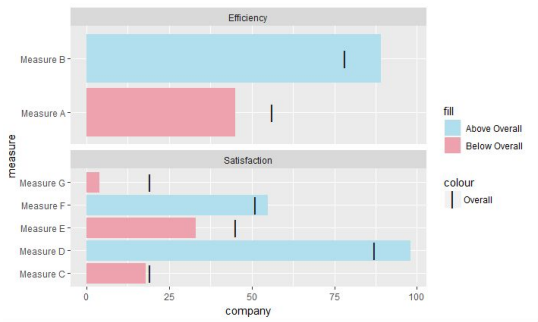
- Including unnecessary elements
- Misusing the Dual-Axis
- Not labeling axes or data points:
- Failing to consider the audience
- Poor color choices
- Including too many variables
- Different length from barplot
- ……











Tools and Databases Related to Data Visualization

Awesome-genome-visualization

https://cmdcolin.github.io/awesome-genome-visualization/

- List of interesting genome visualizers, genome browsers, or genome-browser-like implementations
- Filter for tools with different tags (ex. CNV, Cancer, Gene fusion, Heatmap), programming languages (ex. R, Python C++), and platforms (ex. Interactive, Mac, Web-based)
- Sort tools by recently added, number of citations, etc.

pixy

https://pixy.readthedocs.io/en/latest/plotting.html#a-genome-wide-plot-of-summary-statistics

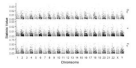
Language: Python, R, ggplot2

Tags: Population, Variation

Note: The link in readthedocs shows a nice general purpose way to plot multi-chromosome plots in ggplot2 with facet grid. Alternative methods for multi-chromosome plots shown by the manhattan ggplot2 tutorial (https://danielroelfs.com/blog/how-i-create-manhattan-plots-using-ggplot/ uses cumulative bp instead of facet_grid)

Github: https://github.com/ksamuk/pixy/

Github Stargazers: 89



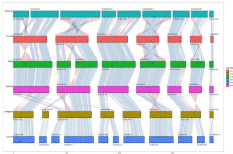
chromsyn

Language: R

Tags: Comparative, Synteny, Multi-way

Github: https://github.com/slimsuite/chromsyn

Github Stargazers: 9



NGenomeSyn

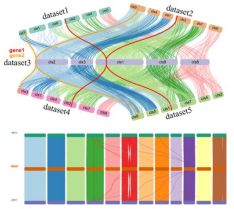
Publication: (doi link) (2023) (# citations 0)

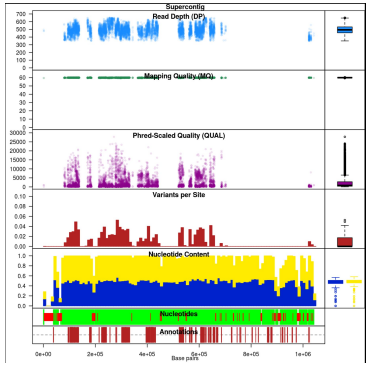
Language: Perl

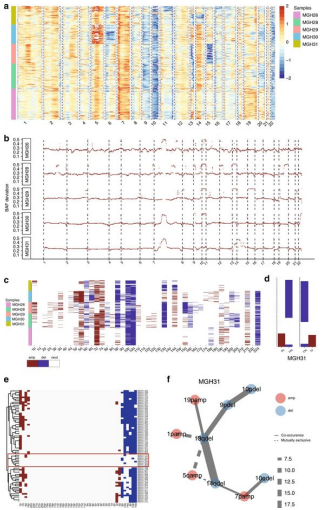
Note: See also RectChr

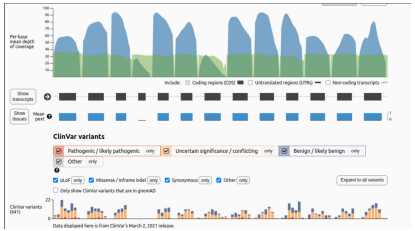
Github: https://github.com/hewm2008/NGenomeSyn

Github Stargazers: 80









Circos Plots

- Display complex, multidimensional data
- Ideal for genomics and bioinformatics
- Display multiple types of data in a single plot
- Highly customizable
- Reveal patterns and relationships
- Useful for exploratory data analysis
- Communicate research findings effectively

http://circos.ca/

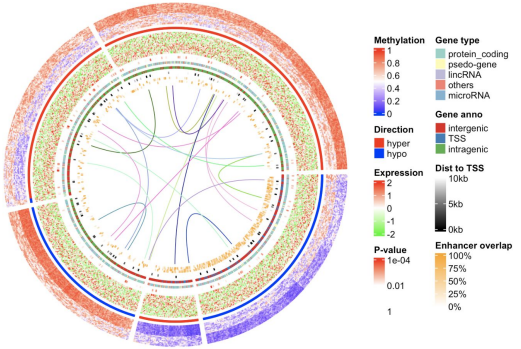
Bioinformatics · Genome Biology · nature · Science · Nucleic Acids Research · American Scientist · Genome Research · Conde Nast Portfolio · WIRED · PNAS · PLoS Public Library of Science · The Plant Cell · The New York Times · Leukemia · SEED

My images created with Circos have appeared in a variety of publications. Wired, New York Times, Conde Nast Portfolio, and American Scientist.

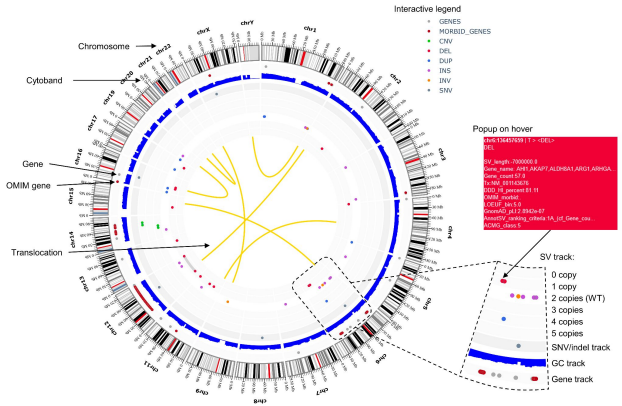
In genomics, scientific journals like Science, Nature, PLoS, Genome Research and others have published papers that used Circos images (Circos citations).

R - Circlize

https://github.com/jokergoo/circlize



Python - vcf2circos

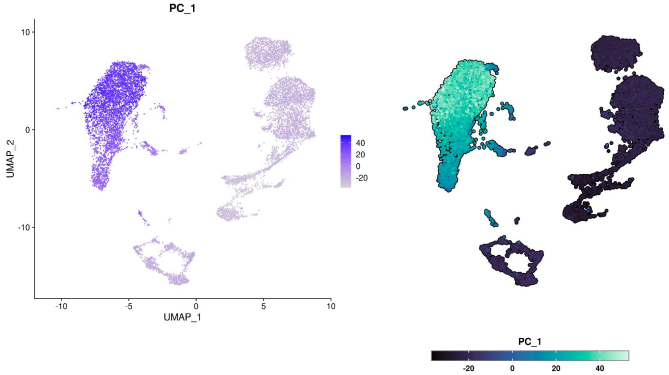


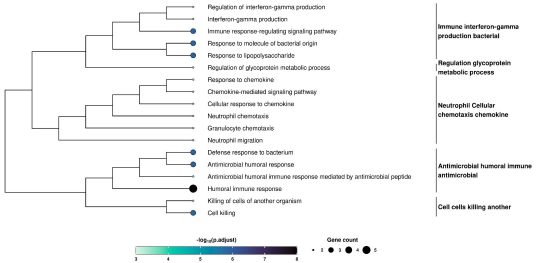
https://github.com/bioinfo-chru-strasbourg/vcf2circos

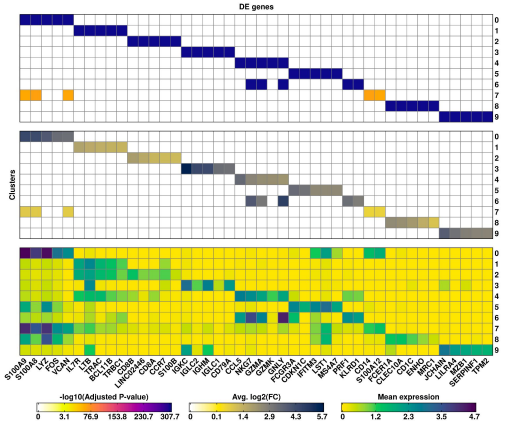
SCpubr

This package aims to provide a streamlined way of generating publication ready plots for Single-Cell transcriptomics in a “publication ready” format (SCpubr). That is, the aim is to generate plots with the highest quality possible, which can be used right away or with minimal modifications for a research article.

https://github.com/enblacar/SCpubr/

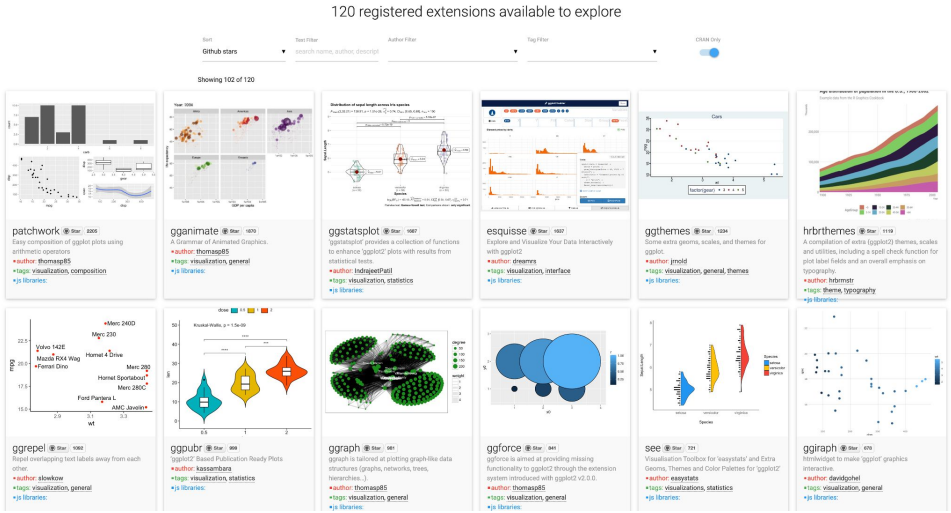


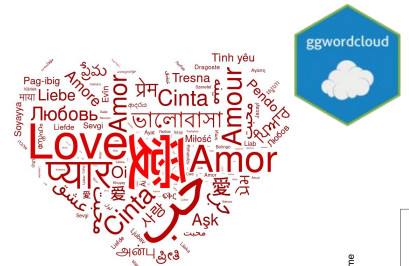


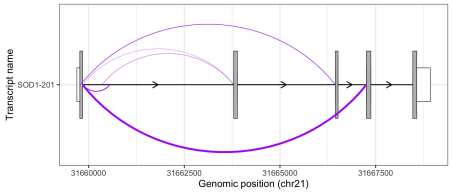


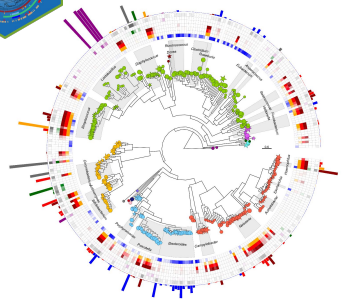
Beyond the basics: exploring the power of ggplot2's extension packages

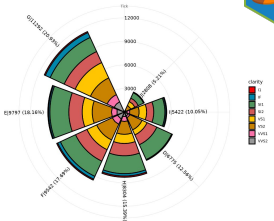
https://exts.ggplot2.tidyverse.org/gallery/











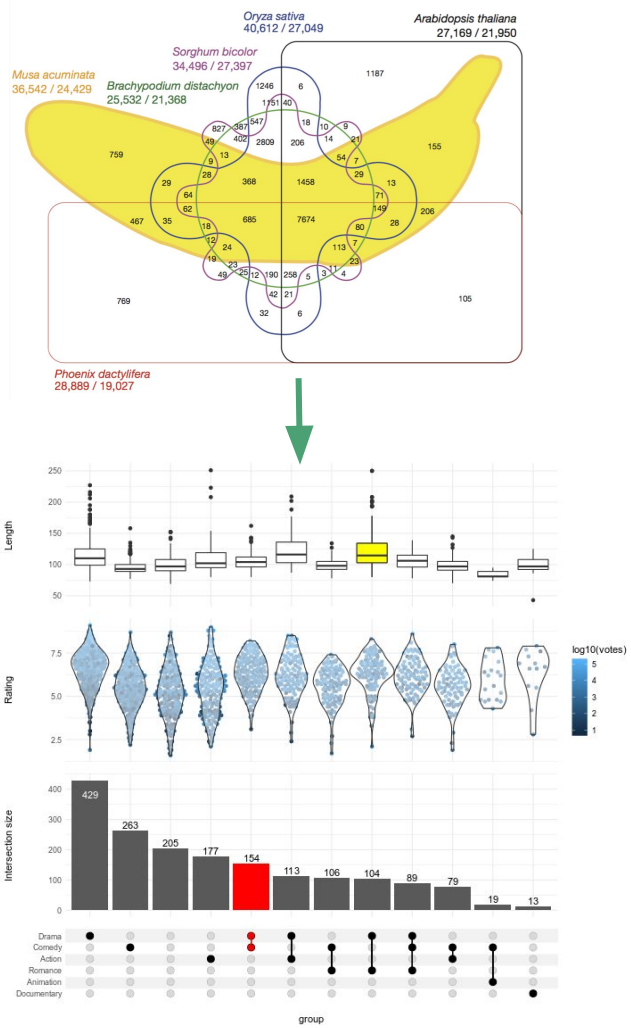
Visualizing intersecting data by UpSet

- Use **perceptually efficient visual encodings**, i.e., make it easy to read the data accurately.
- Make it possible to not just visualize intersections, but to **visualize combinations of intersections** (e.g., all the intersections involving two particular sets).
- **Visualize attributes about the intersections**. It is not just the magnitude of an intersection that is interesting, but we also want to know whether the data associated with intersection is different or similar.

App:
https://upset.app/

R package: UpSetR
https://github.com/hms-dbmi/UpSetR

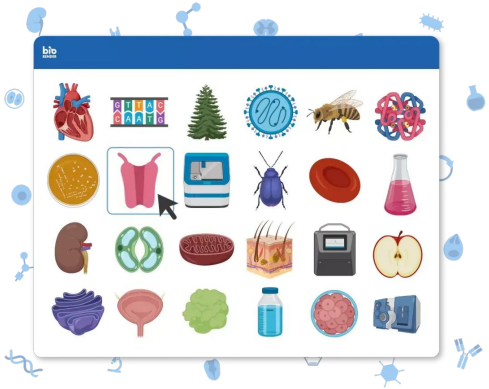
Python Package: UpSetPlot
https://github.com/jnothman/UpSetPlot



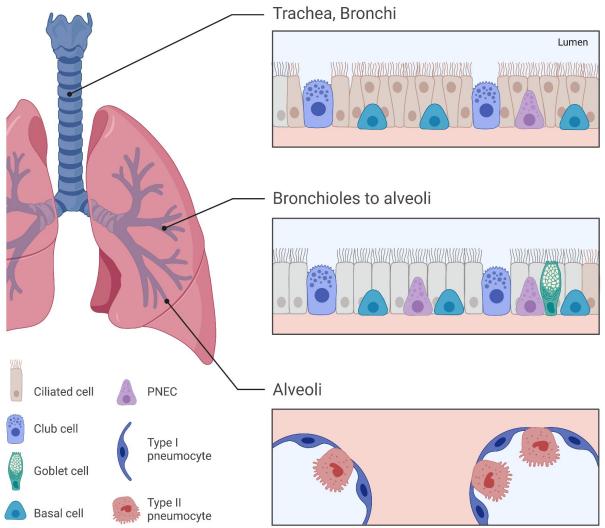
UpSet

BioRender: Scientific Image and Illustration Software

- Customize pre-made illustrations to fit user needs
- Simple drag-and-drop functionality
- Work together in real-time with team members
- Access to a library of scientific icons: Choose from 50,000+ icons and templates
- Templates for scientific posters, figures, professional-looking visuals for publication



https://app.biorender.com/portal/nci



Used by thousands of trusted institutions:

JOHNS HOPKINS MEDICINE · UNIVERSITY OF CAMBRIDGE · Roche · CellPress · UNIVERSITY OF TORONTO

Genentech · Stanford MEDICINE · PRINCETON UNIVERSITY · abbvie · NIH National Institutes of Health

Understanding Cancer Genetics Data and Creating Effective Visualizations

Visualizing complex data

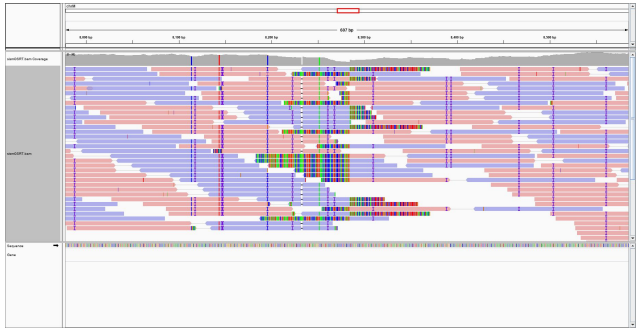
The problem has become more acute with the ever-increasing amount and complexity of scientific data. Visualization of those data — to understand as well as to share them — is more important than ever.

1. Select the appropriate visualization technique for the type of data you are working with.
2. Highlight key features using color, size, or other visual cues to make them stand out.
3. Simplify the visualization to avoid overwhelming the audience with too much detail.
4. Use interactive elements to allow the audience to explore the data in their own way.
5. Provide context by including background information and definitions to help the audience interpret the data.
6. Ensure that the visualization is accessible to the audience by using clear labels and titles that are easy to understand.

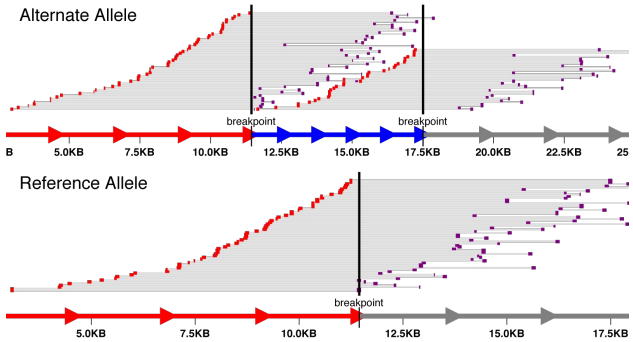
Visualizing raw sequencing data

IGV GUI to Command Line

https://github.com/hartleys/igvSnap



https://github.com/svviz/svviz



BamSnap

pypi v0.2.19 | downloads 154/month | docs passing | docker pulls 1.6k

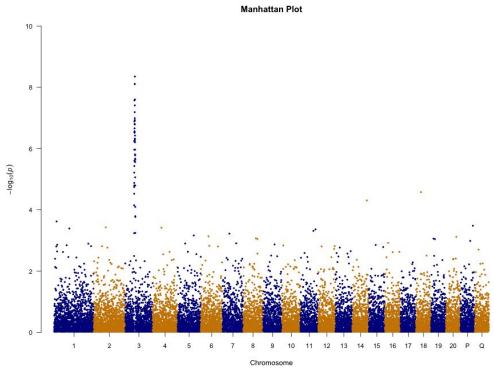
BamSnap is a visualization tool for aligned BAM files that allows to generate high-quality snapshots of read level data in high-throughput, processing up to thousands of files. **BamSnap** is a command-line software based on python.

https://github.com/parklab/bamsnap

Data visualization with genomic coordinates

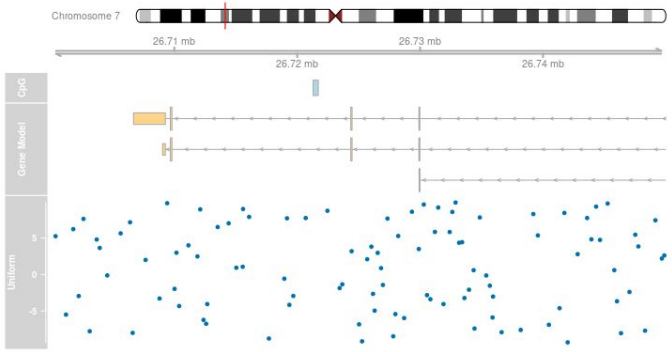
Manhattan Plot

https://github.com/stephenturner/qqman



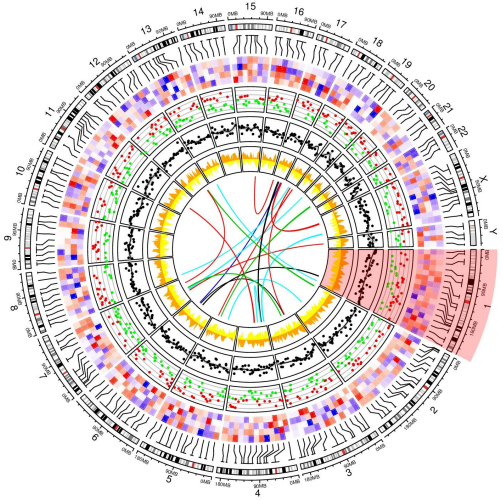
Gviz Plot

https://bioconductor.org/packages/release/bioc/html/Gviz.html



Circos Plot

https://github.com/jokergoo/circlize

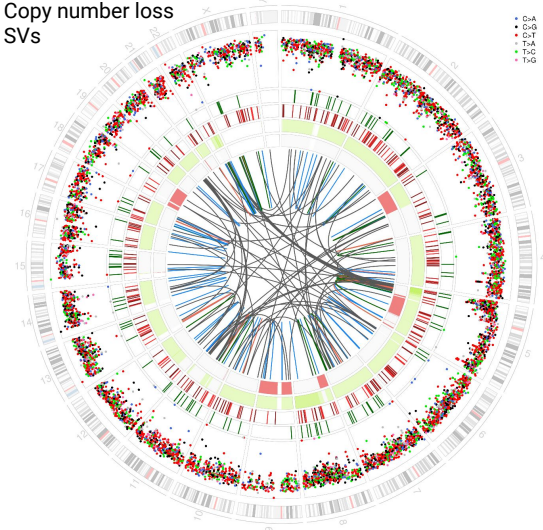


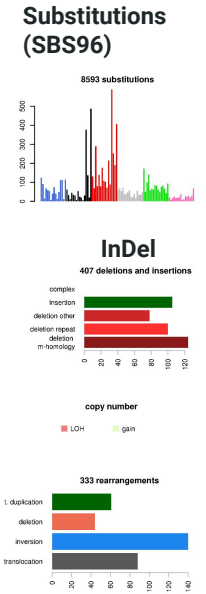
Data visualization for cancer genomes (sample level)

genomePlot

Tracks (outer -> inner):

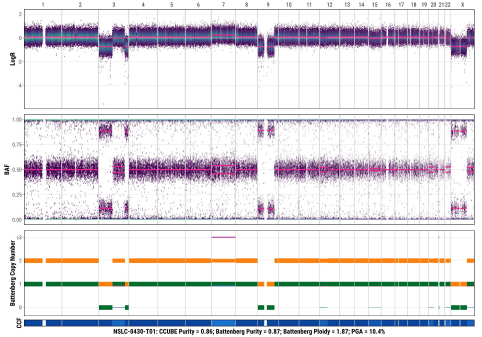
- SNVs
- insertions
- Deletions
- Copy number gain
- Copy number loss
- SVs

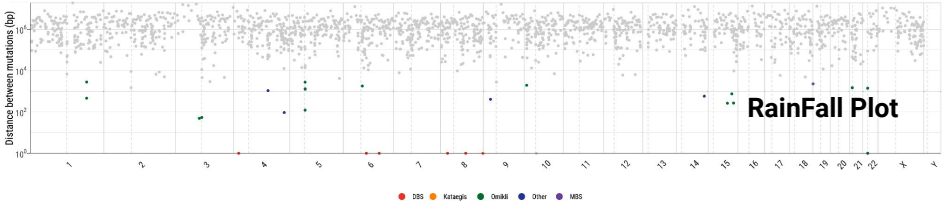




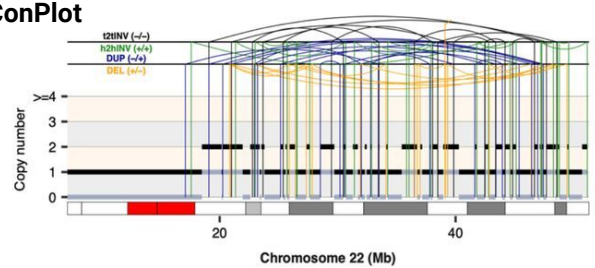
signature.tools.lib R package:
https://github.com/Nik-Zainal-Group/signature.tools.lib

SCNA plot





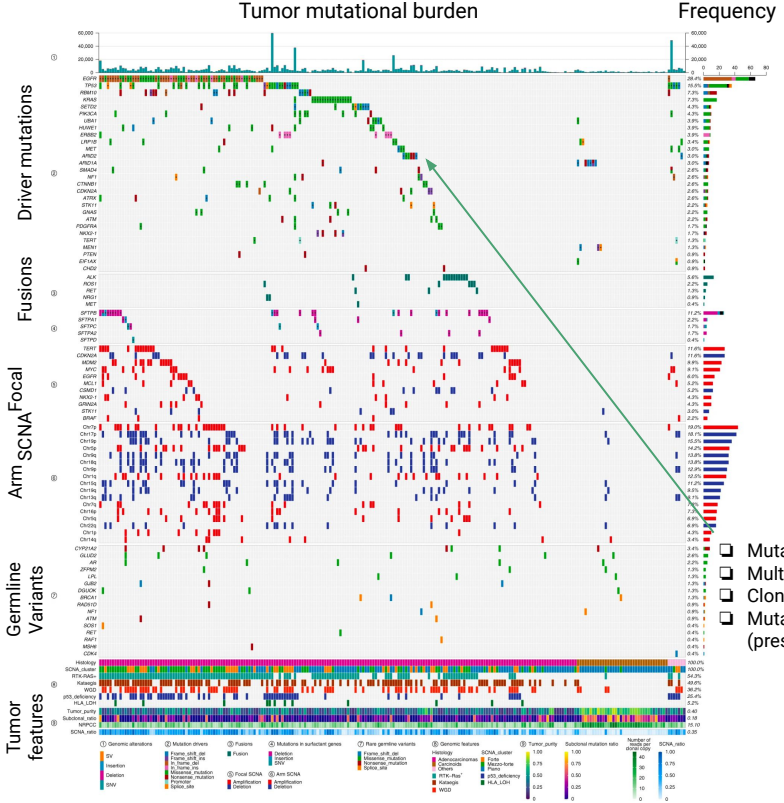
ReConPlot



https://github.com/cortes-ciriano-lab/**ReConPlot**

Data visualization for genomic features (study level)

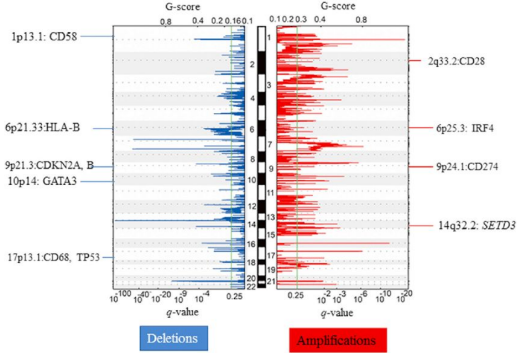
Genomic landscape - Oncoplot



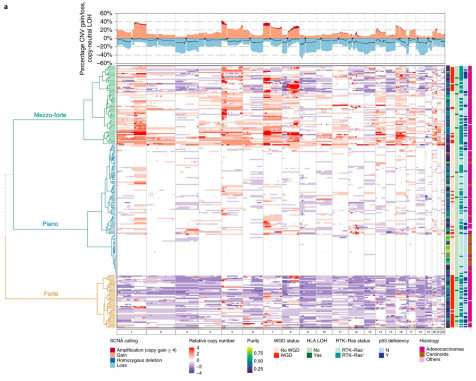
- Mutation effect (color)
- Multiple mutations (size)
- Clonality (circle)
- Mutational relationship (present vs absent)

https://github.com/PoisonAlien/maftools

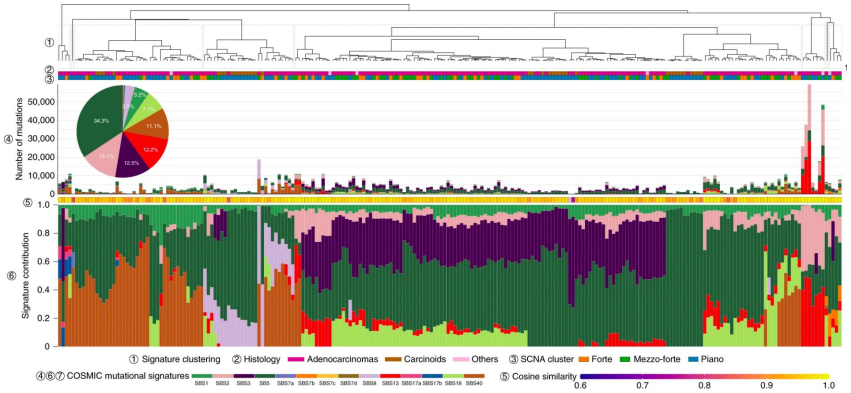
Significant focal SCNA events



SCNA classification



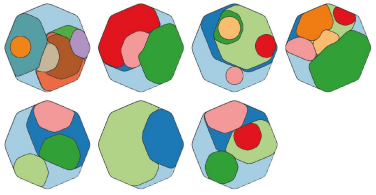
Mutational processes landscape

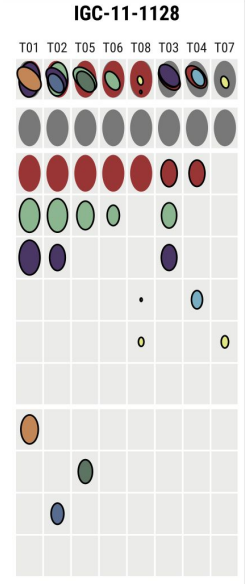


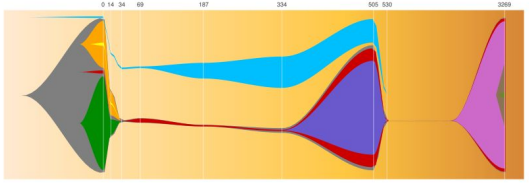
Data visualization for tumor evolution

Clonal architecture

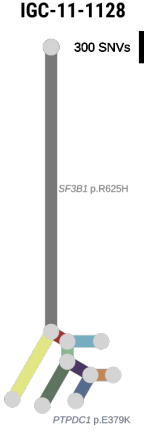
https://github.com/amf71/cloneMap





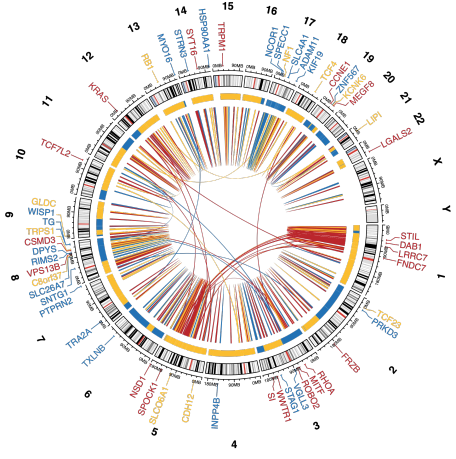


https://github.com/chrisamiller/fishplot

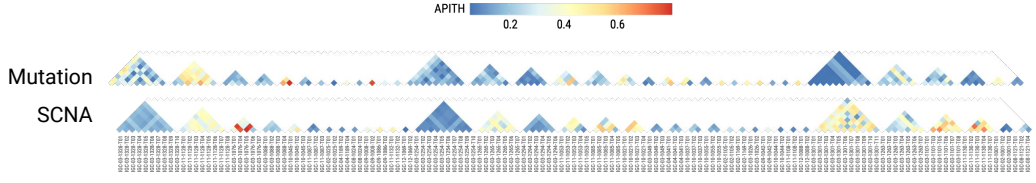


cloneTree: under development

Intratumor Heterogeneity (ITH)

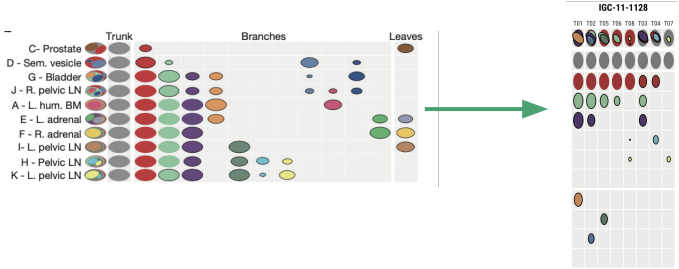


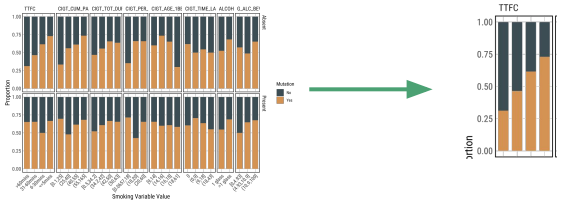
Blue: trunk
Yellow: branches
Red: Leaves

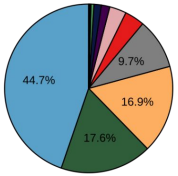


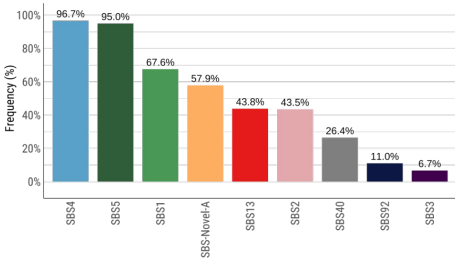
Suggestions for Creating Publication-Ready Figures

- ★ Choose the right visualization technique for your data, considering both the type of data and the message you want to convey. (you may get some ideas from other visualizations with similar data types or analysis)
- ★ Keep it simple and clear: Focus on the most important message you want to convey and remove any extraneous details or clutter that could confuse the audience. (Especially for the main figures in a manuscript)
- ★ Use color effectively: Choose a limited color palette and use color to highlight important information or patterns in the data. Avoid using too many colors or making color choices that may not be visible for people with color vision deficiencies. (Use consistent colors across all figures in the manuscript)



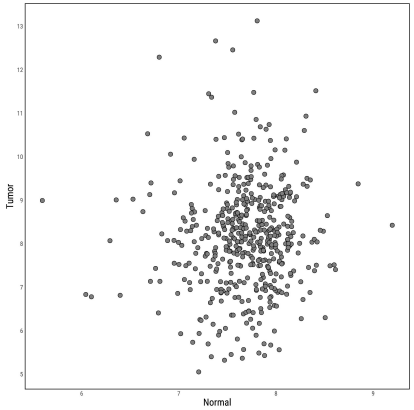


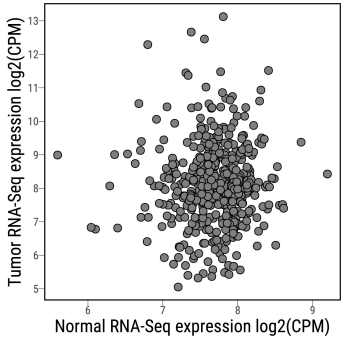




★ Label everything clearly: Provide clear and concise labels for all axes, legends, and data points to help the audience understand the figure. Use a legible font size and style that can be easily read, even when the figure is reduced in size for publication. (Keep in mind the minimal font size in the very beginning.)

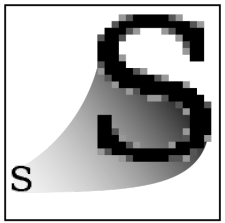
★ Consult the journal's guidelines: Review the journal's guidelines for figures to ensure that your figure meets their requirements for size, resolution, and format. (Try to generate smaller figures with a clear message)





- ★ Ensure reproducibility: Provide all necessary details about the methods used to create the figure, including any software or code used. This will help other researchers to reproduce your results or build upon your work. (For R, save the entire image object for future investigation and reproduction)
- ★ Choose the right format: Select a format that will be suitable for publication, such as high-resolution image files or **vector graphics**. Consult the journal's guidelines or requirements to ensure you're using an appropriate format. (PDF, SVG, or EPS are all good vector graphics).
- ★ Get feedback: Ask colleagues or collaborators to review the figure before submitting it for publication to get feedback on its clarity and effectiveness.

 # For reproduce figures
>save.image (file = 'xxx')

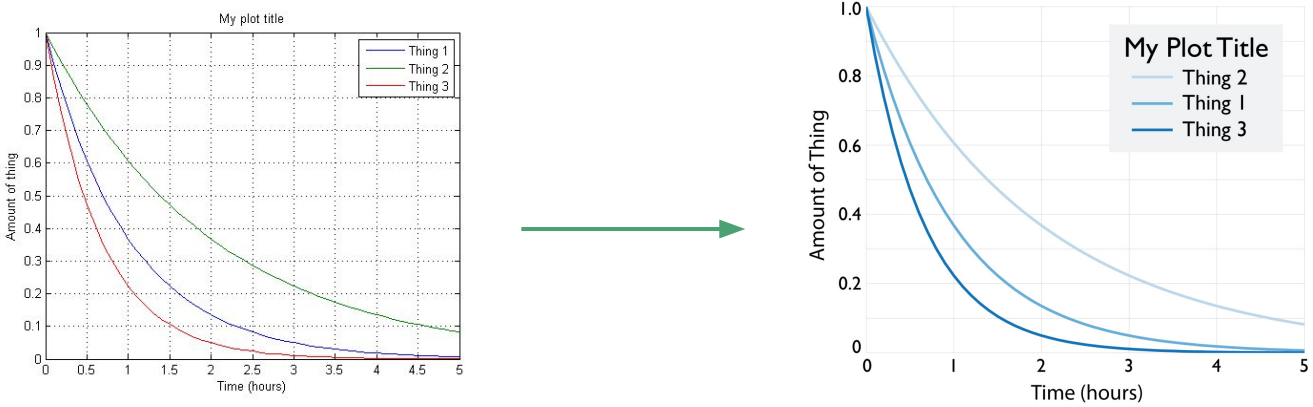


Editing Scientific figures using Adobe Illustrator ('AI')

Over figure-scripting?

Because it's a lot of trouble to fully script figures, and hardcoding everything can make it very difficult to make requested edits. Adobe Illustrator can:

- Resize images/figures
- Add scale bars
- Easily modify colors and rearrange panels
- Easily adjust aesthetics (e.g. alignment, compiling micrographs with schematics)

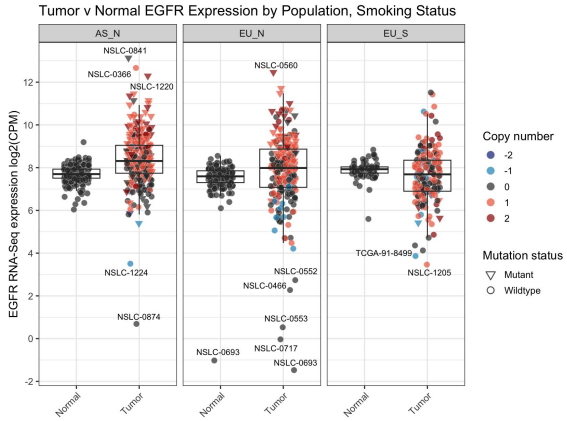


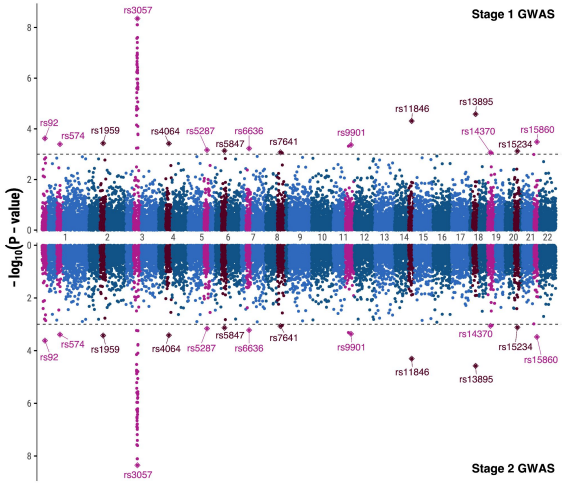
basic Adobe Illustrator (CC) guide

THANKS FOR YOUR ATTENTION!
Questions?

Next: Practical session 12 (10:45 am)

- ❏ Practice data visualization using R ggplot2 package





Date: Thursday, May 25th, 2023

Time: 10:30 AM – 11:30 AM

Speaker: Charles Swanton, MBPh.D., FRCP, FMedSci, FAACR, FRS, Francis Crick Institute

Title: Mechanism of Action and Inflammatory Axis for Air Pollution Induced Non-Small Cell Lung Cancer

Date: Thursday, June 8th, 2023

Time: 10:30 AM – 11:30 AM

Speaker: David Adams, Ph.D., Wellcome Sanger Institute

Title: Cross-species oncogenomics of melanoma and other malignancies to define disease drivers

Date: Thursday, July 13th, 2023

Time: 10:30 AM – 11:30 AM

Speaker: Jinghui Zhang, Ph.D., St. Jude Children's Research Hospital- Department of Computational Biology

Title: Therapy-Related Clonal Evolution in Pediatric Cancer Patients and Long-term Survivors