The Elizabeth H.
and James S. McDonnell III
McDonnell Genome Institute
at Washington University

Genomic data visualization and interpretation

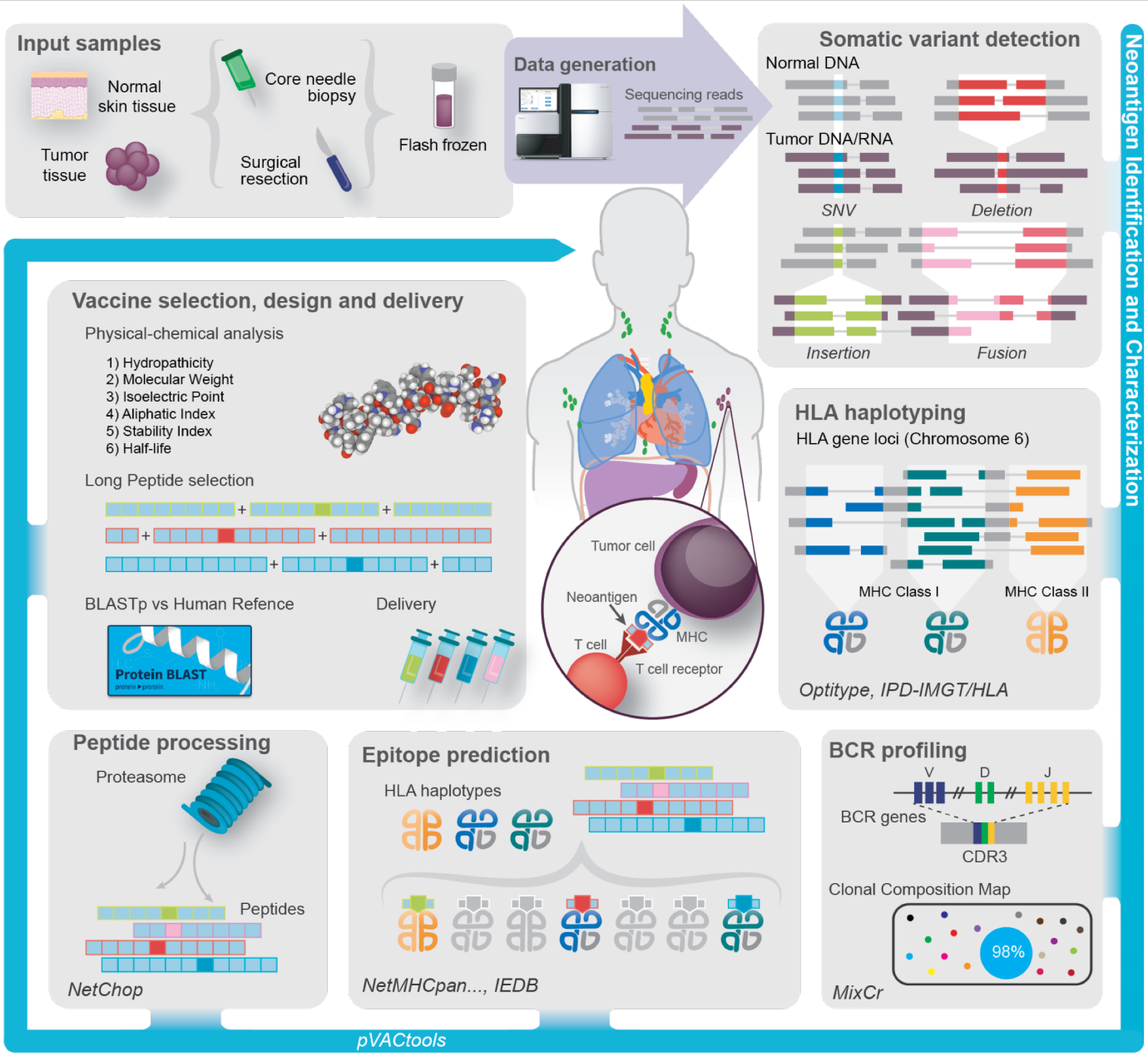
Malachi Griffith, Zachary Skidmore, Obi Griffith

Workshop on Genomics

May 27, 2022

Český Krumlov, Czech Republic

My background: Translational genomics and personalized oncology



griffithlab.org

Why do we create visualizations of genomic data?

- Data exploration and interpretation of results
  - QC analysis
  - Understanding whether/how an experiment worked
  - **Discovery**

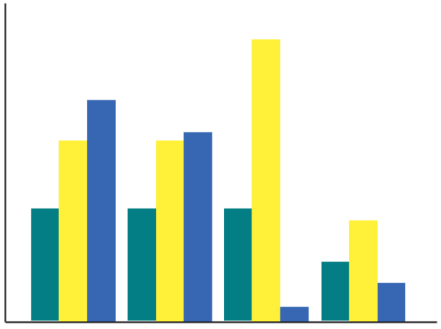
- Communication
  - Slides for presentations
  - Figures for publications

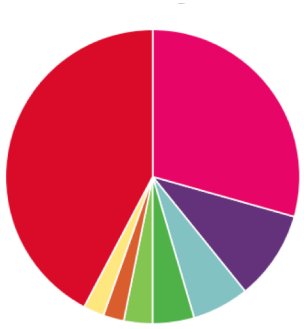
Fundamentals of data visualization

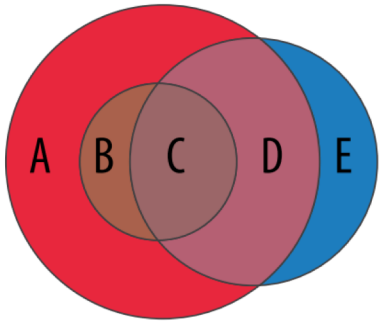
Fundamentals of data visualization

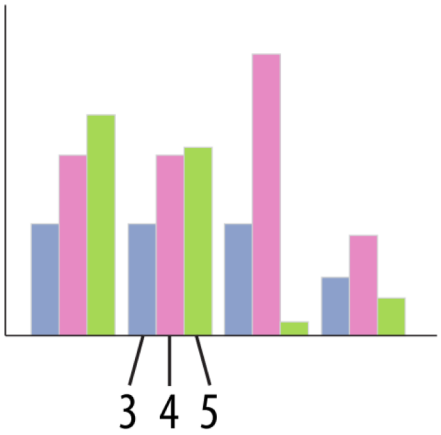
- Where to learn more about the art and science of visualization:
  - Collection of 40 Nature Methods articles on data visualization: "Points of View" Articles
  - Visual design principles lecture

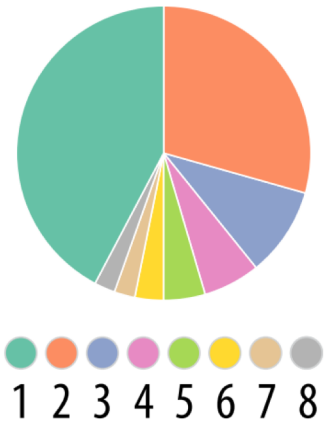
Which series is more effective (top or bottom)?

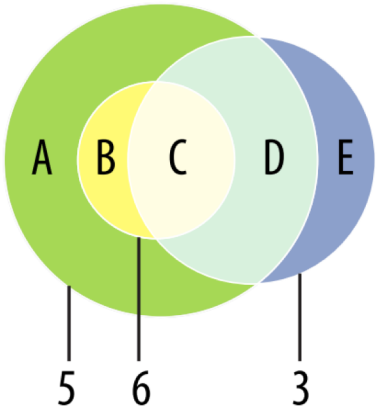








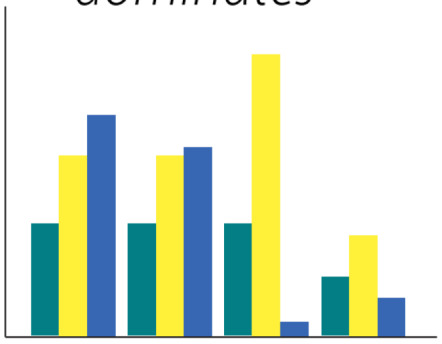




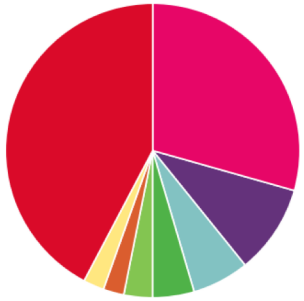
Martin Krzywinski

Which series is more effective (top or bottom)?

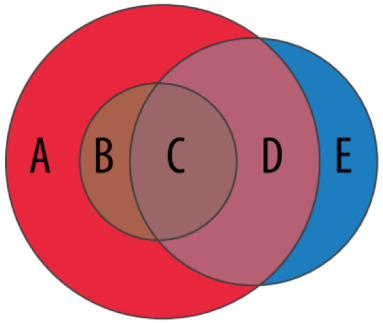
one color dominates



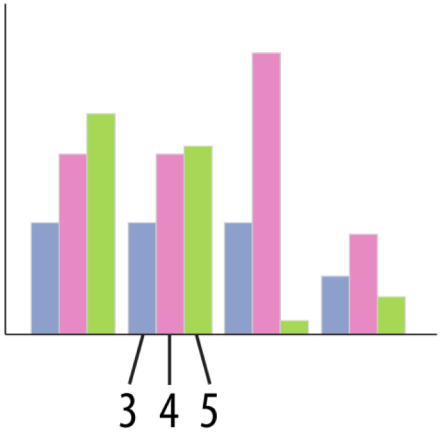
difficult to distinguish

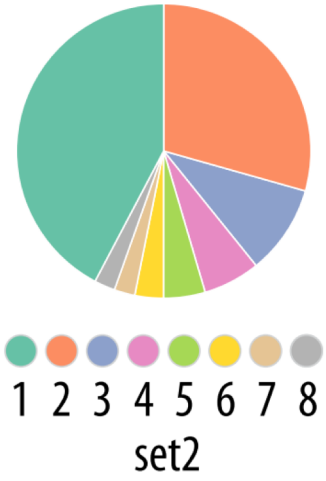


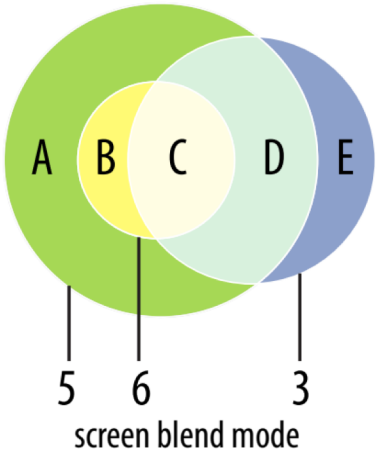
murky



recolored with Brewer palettes

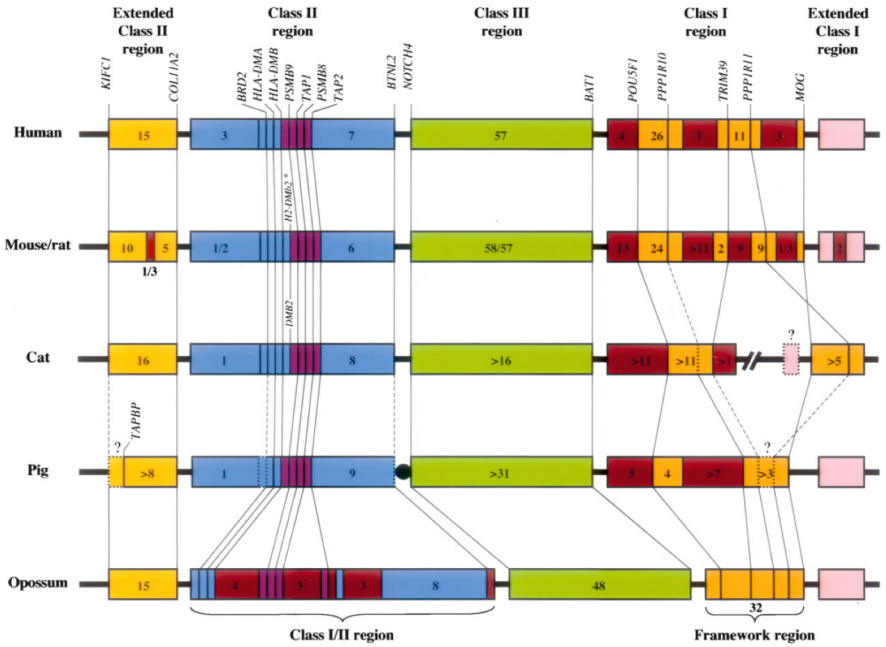


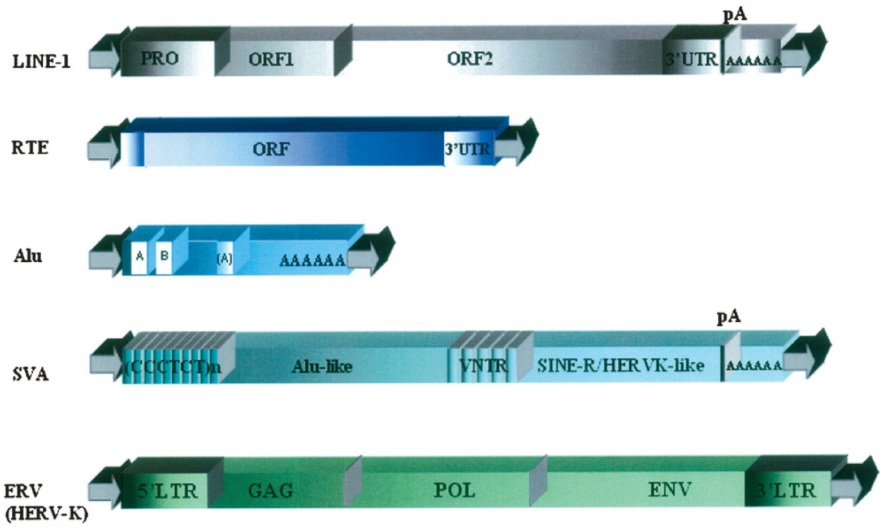




Martin Krzywinski

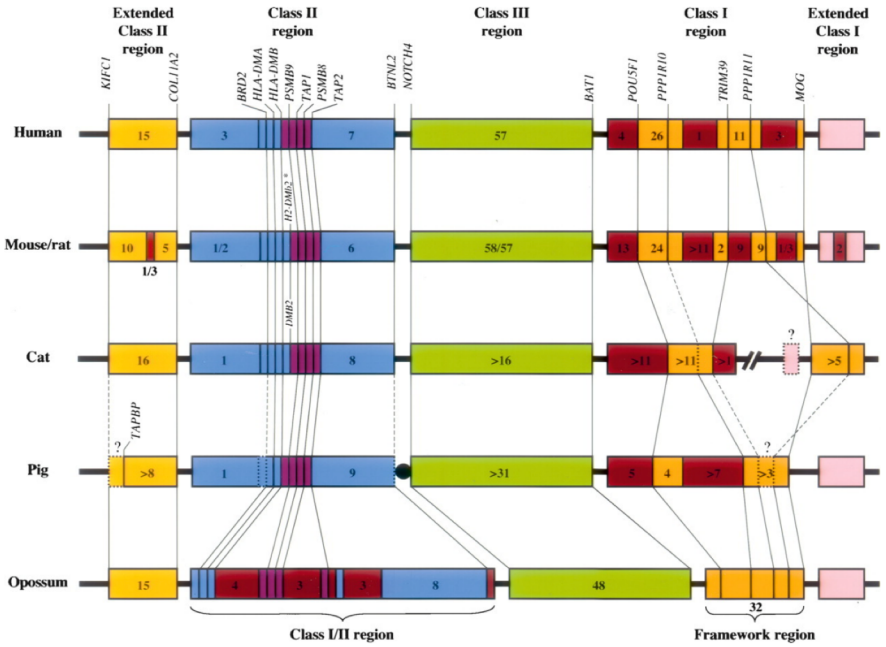
Which is more effective (left or right?)





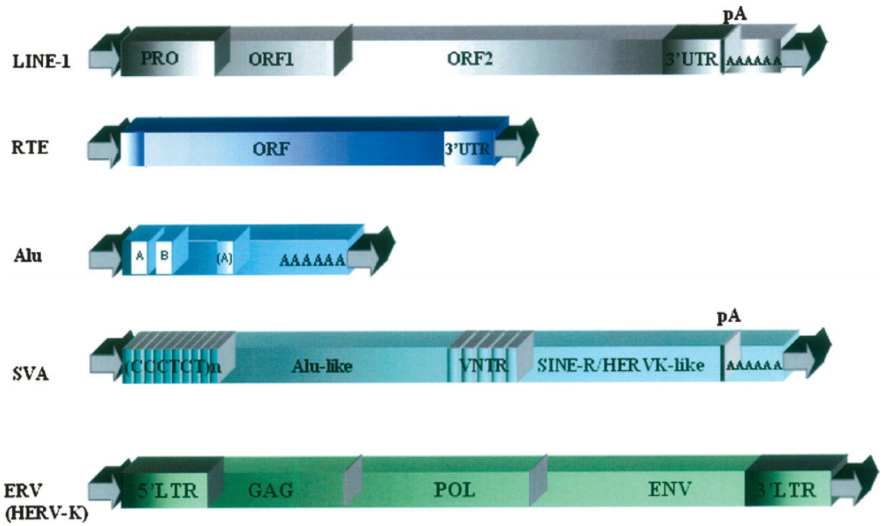
Martin Krzywinski

Which is more effective (left or right?)



Excellent organization and consistency. Vertical lines cue continuity. Good use of color.

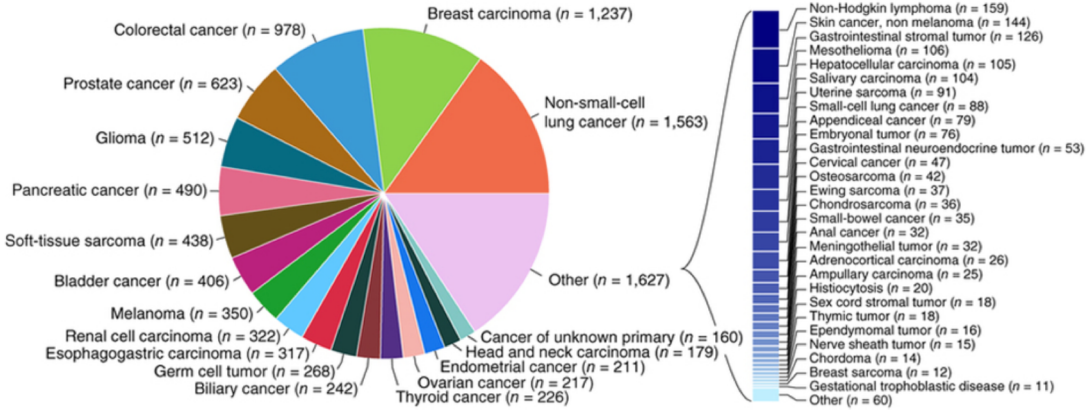
Samollow, P.B., The opossum genome: insights and opportunities from an alternative mammal. Genome Res, 2008. 18(8): p. 1199-215.

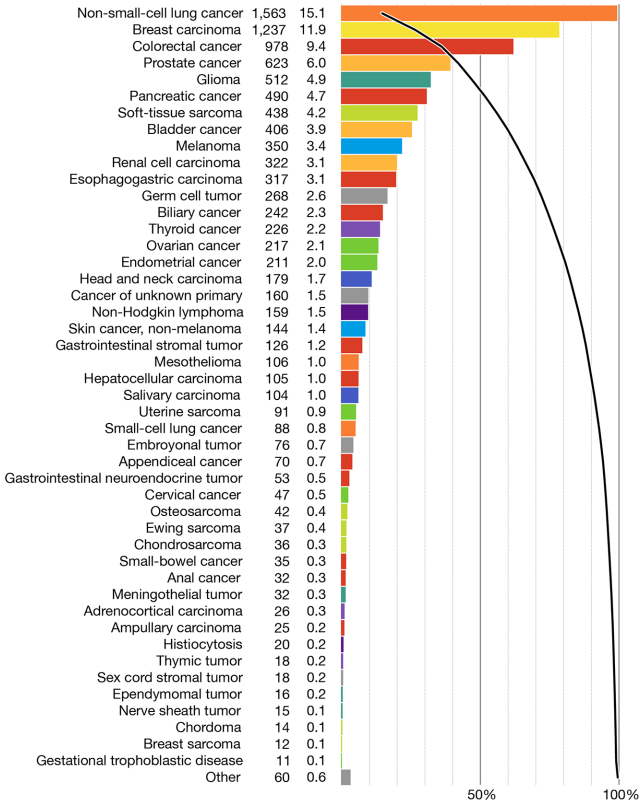


Chartjunk plentiful. Screaming ornamental and redundant elements. Text inconsistent and illegible.

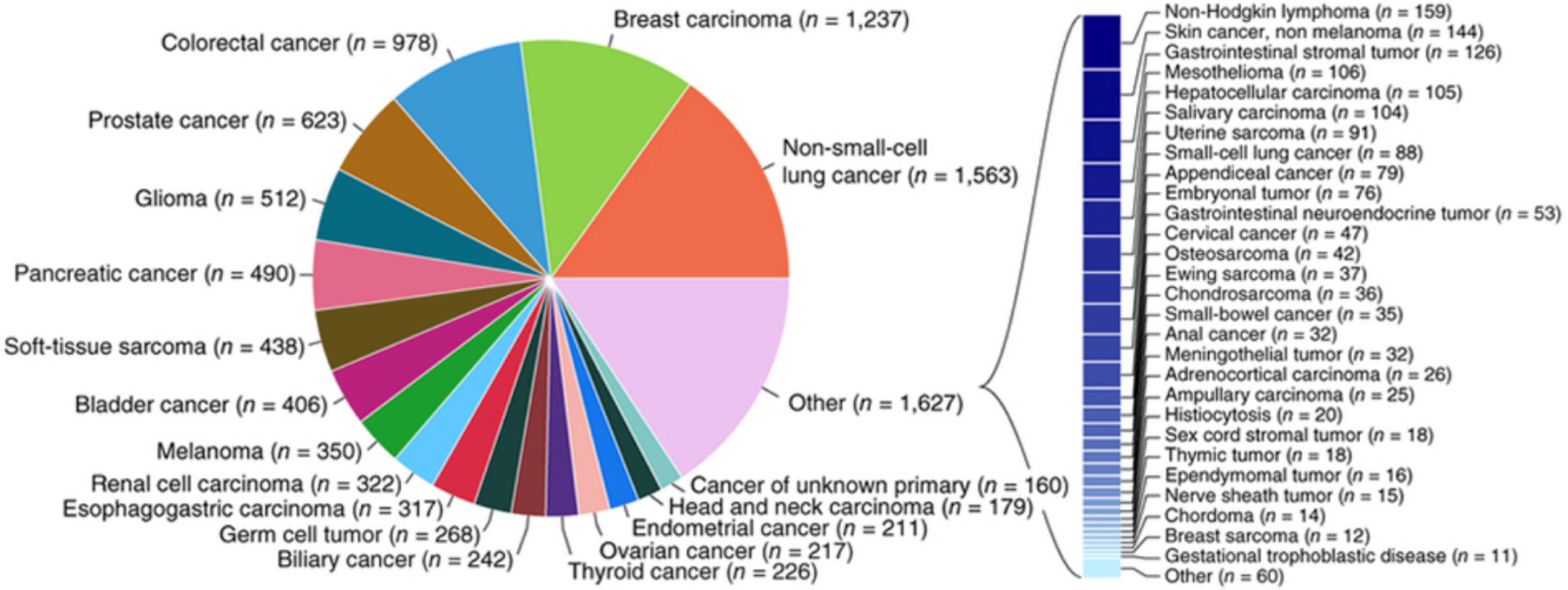
Gentles, A.J., et al., Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. Genome Res, 2007. 17(7): p. 992-1004.

Which is more effective (left or right?)





When to not use a pie chart

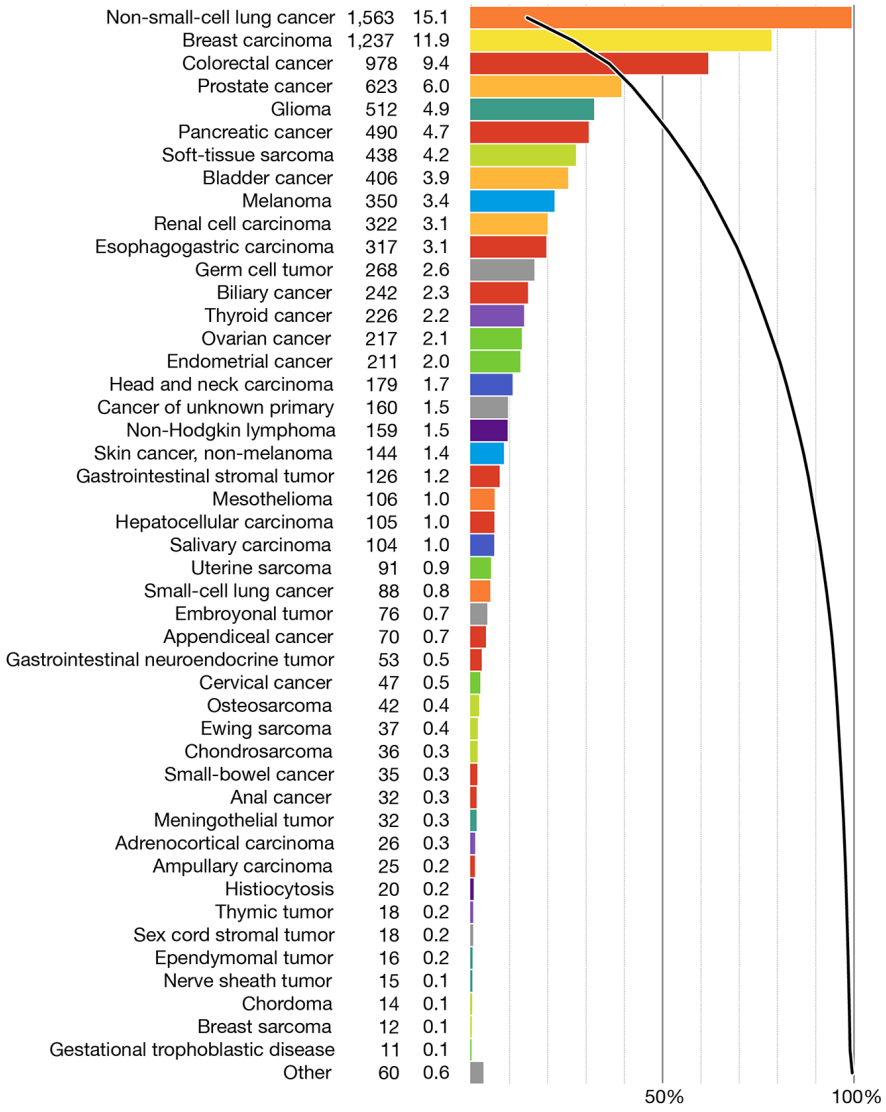


Ahmet Zehir et al. (2017) Nature Medicine doi:10.1038/nm.4333

- Hard to judge proportions
- Poor use of color
- Hard to read labels
- Over ½ of the categories had to broken out of the pie chart

Martin Krzywinski

Same data with a redesigned approach



Martin Krzywinski

When to use a pie chart

Pie charts are good at precisely showing 1:3 proportions

...but not if the slice is rotated

Martin Krzywinski

Selected articles on fundamentals of data viz

- Visualizing samples with box plots
- Circos plots
- When to use (and not use) pie charts

- Resources for choosing colors
  - http://colorbrewer2.org/
  - http://mkweb.bcgsc.ca/color/
  - Understanding and using Color Palettes
  - Color palettes for color blindness
  - Names for >9000 colors
    - Including 40 beer colors

- Credit to Martin Krzywinski for his extensive work in this area and many of the above resources

Best practices in visualization

Best practices from this workshop

- *Always* label axes
- Consider readability of font size
- Avoid vertical or angled text if possible
- Avoid unnecessary use of color, point shapes, etc.
- Chose colors wisely
- If individual data points are being plotted and have started to really pile up on top of each other consider using a density function
- Always be transparent about what data manipulation is taking place (e.g. log scale, filtering of outliers, etc.)

Best practices from the experts

Ten Simple Rules for Better Figures (Rougier et al. 2014):

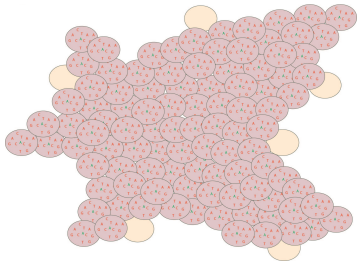
Scientific visualizations should act as a “a graphical interface between people and data”. Try to follow these rules.

1. Know Your Audience
2. Identify Your Message
3. Adapt the Figure to the Support Medium
4. Captions Are Not Optional
5. Do Not Trust the Defaults
6. Use Color Effectively
7. Do Not Mislead the Reader
8. Avoid “Chartjunk”
9. Message Trumps Beauty
10. Get the Right Tool

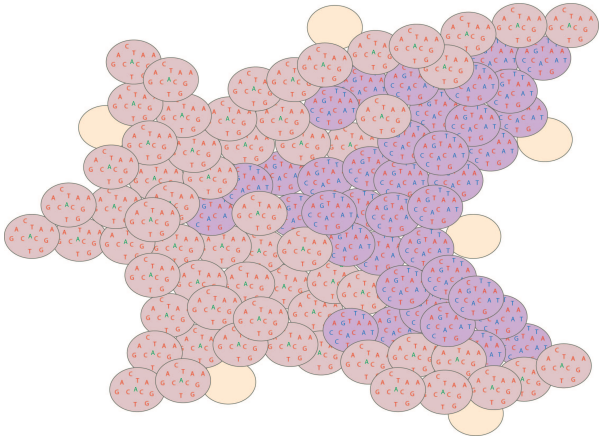
Example visualizations using R

Visualizing tumor evolution

Tumor initiation

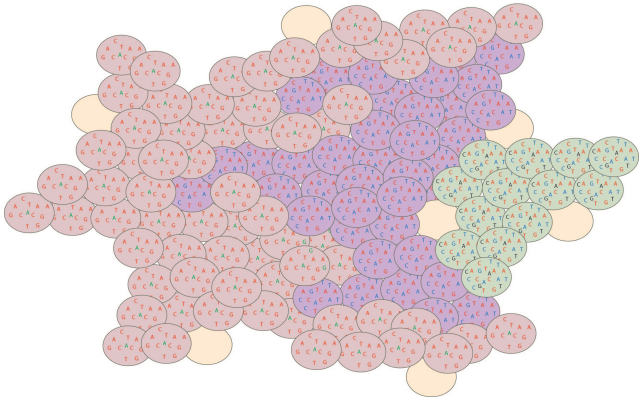


Unknown time

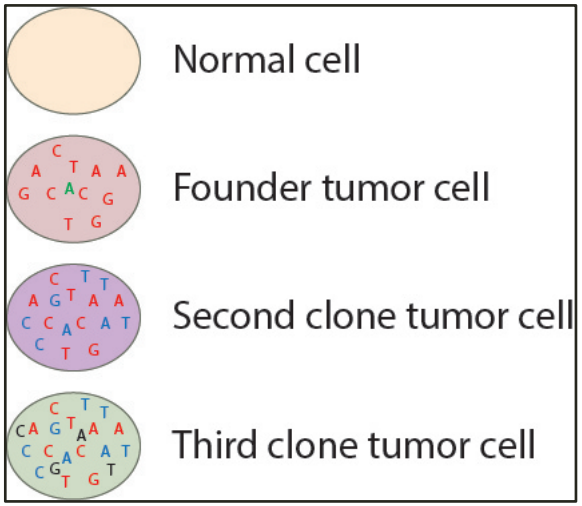


Primary tumor
(at diagnosis)

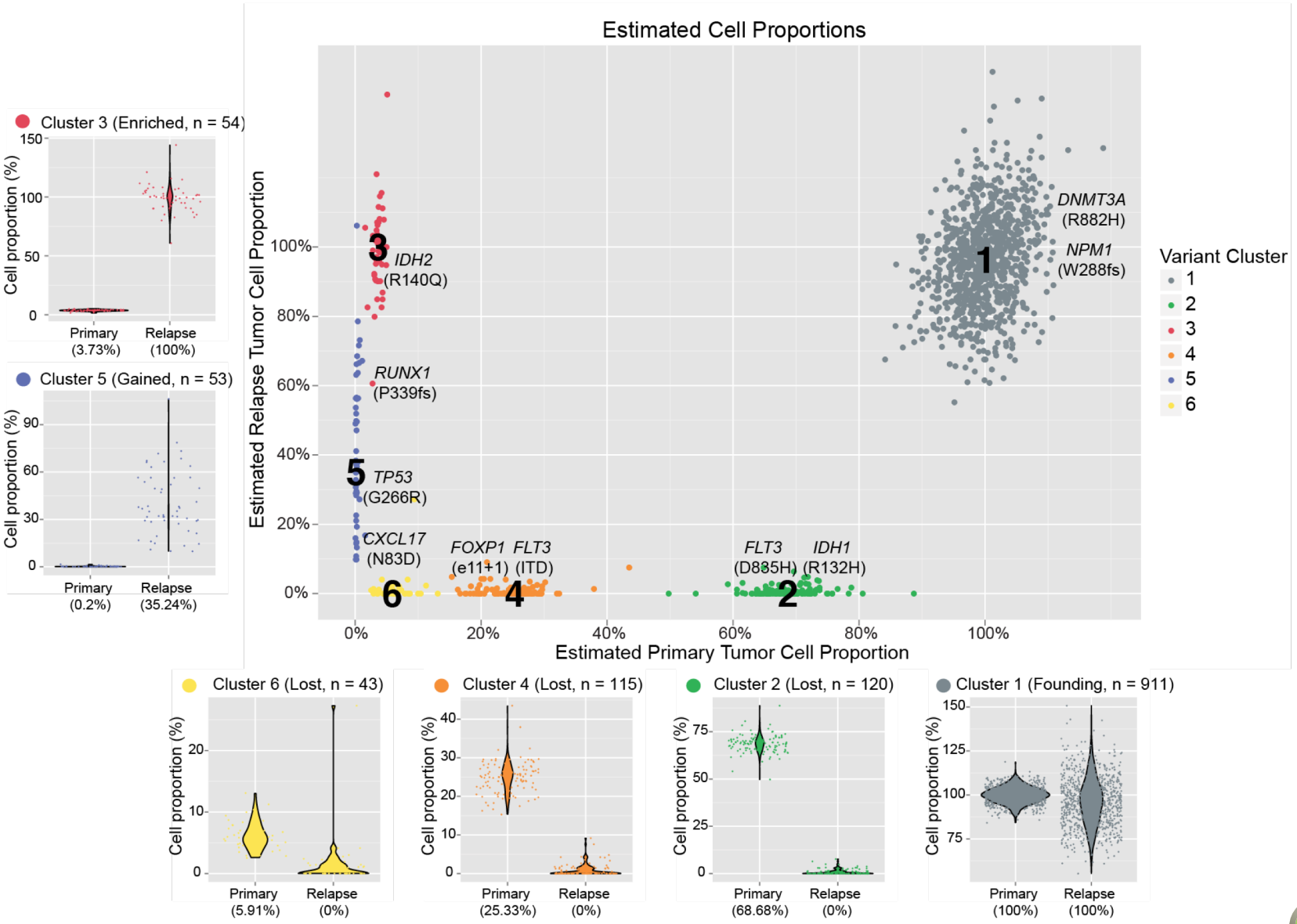
1.5 years



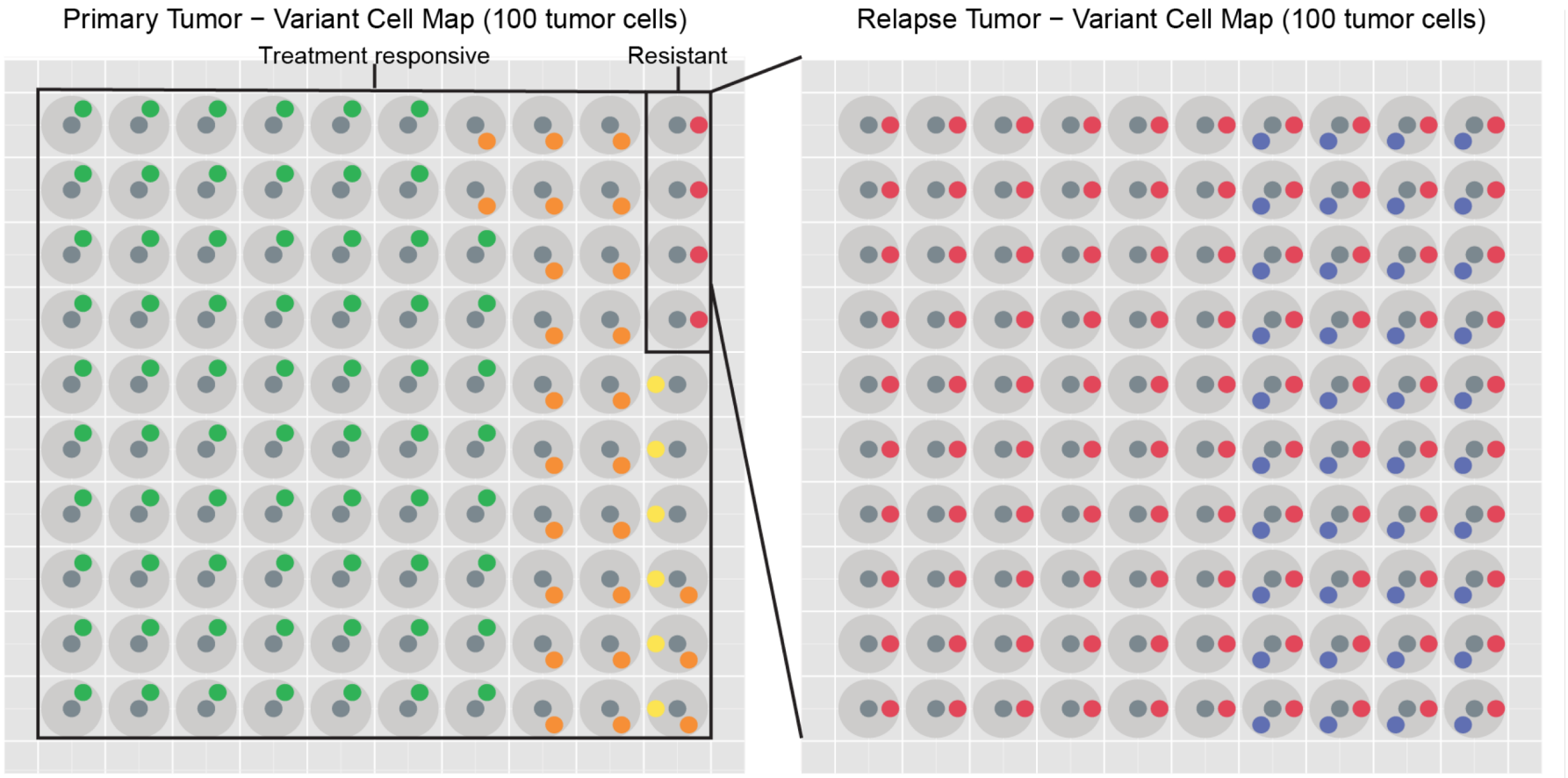
Recurrence



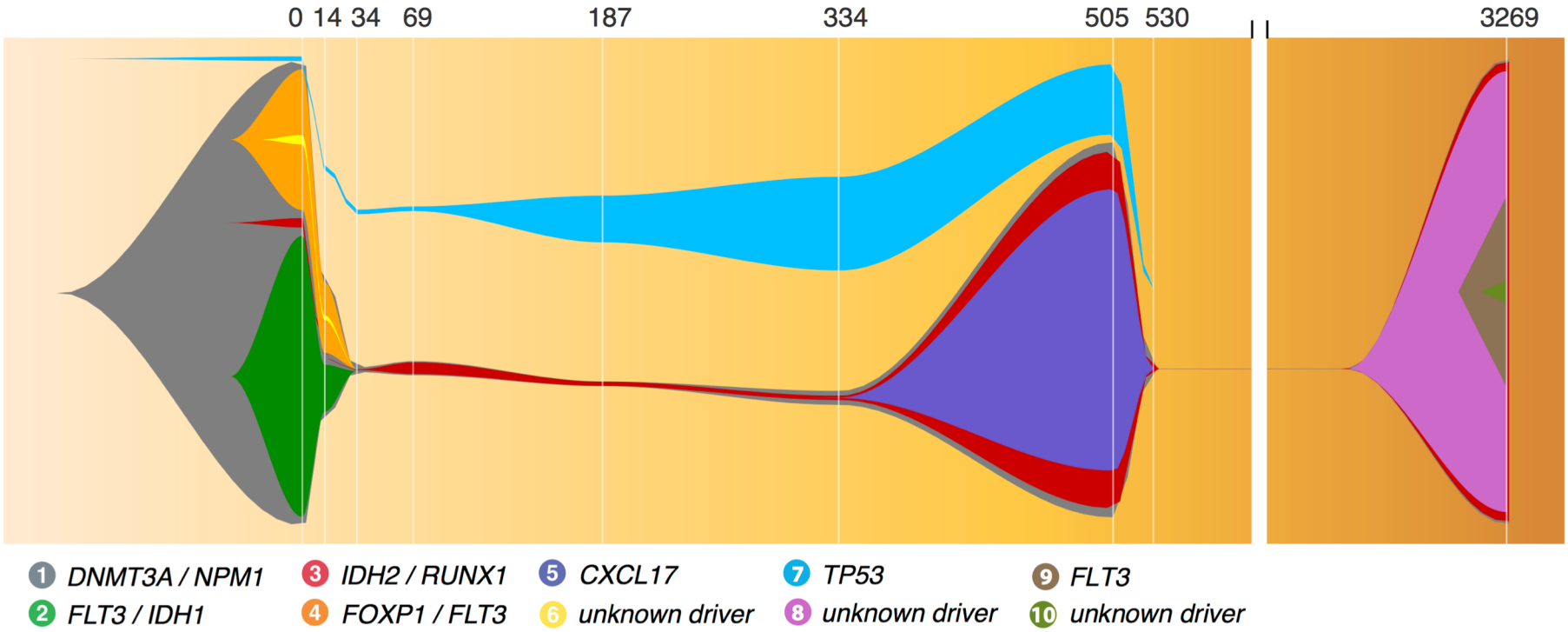
Clustering VAFs to define to distinct cell populations



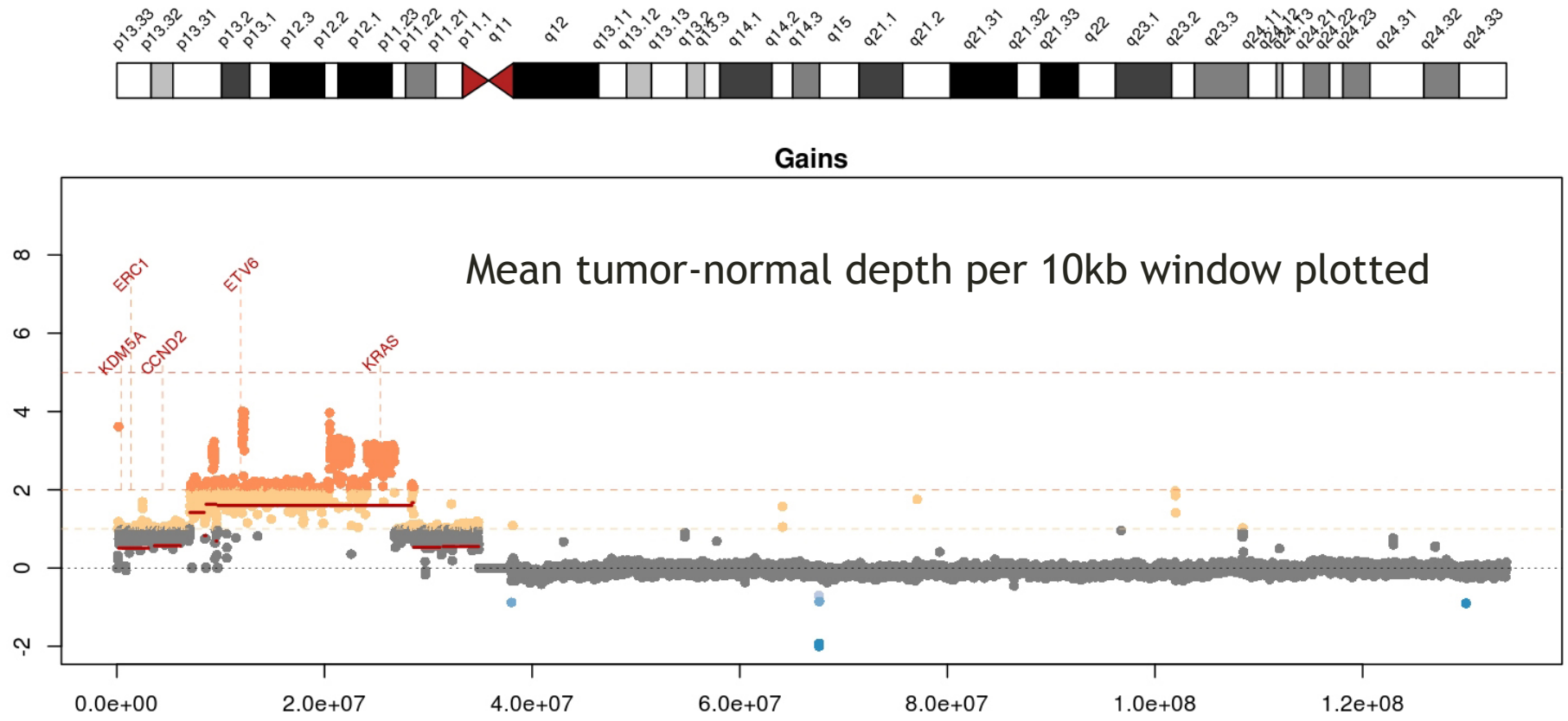
Using a “cell map” to represent the variant clusters in 100 hypothetical cells of a tissue



A ‘fish’ plot is used to represent sub-clones lost and gained over time

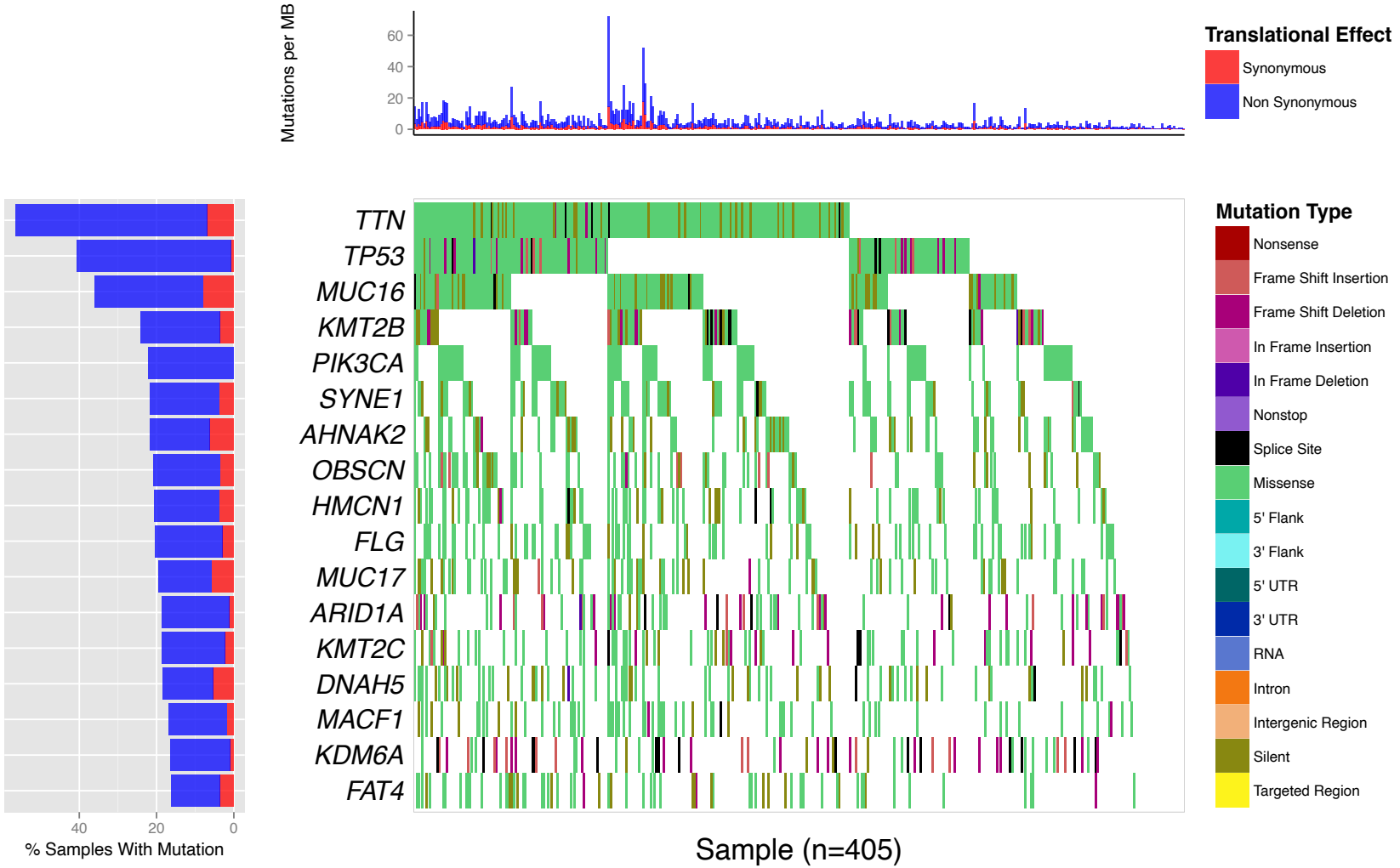


Visualizing copy number variation



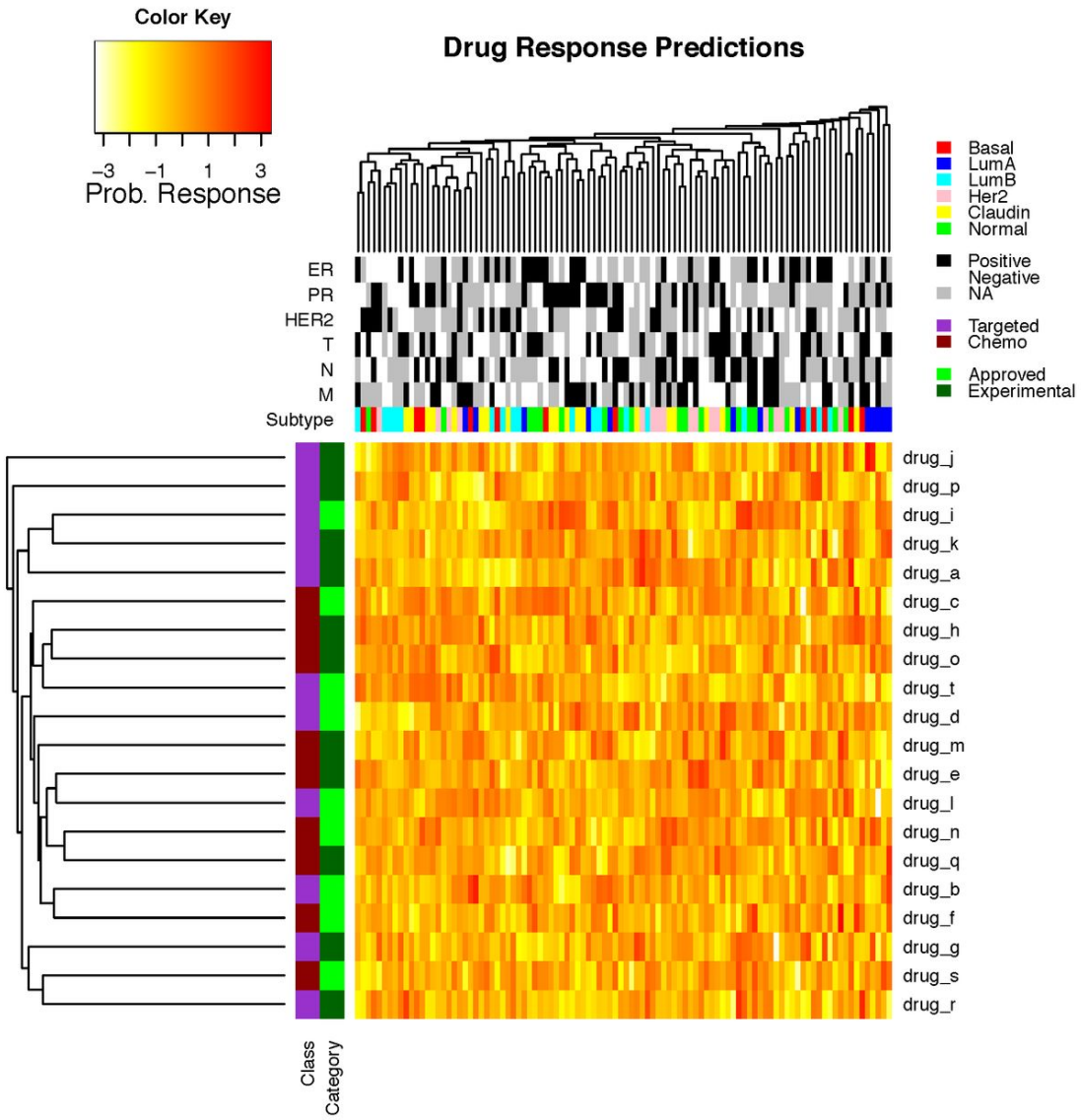
KRAS amplification in a metastatic breast cancer

A ‘waterfall’ plot is one way to visualize the pattern of variant recurrence in a cohort of samples



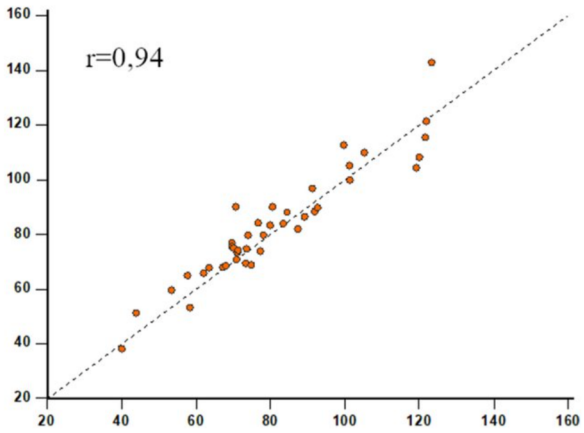
https://github.com/griffithlab/GenVisR

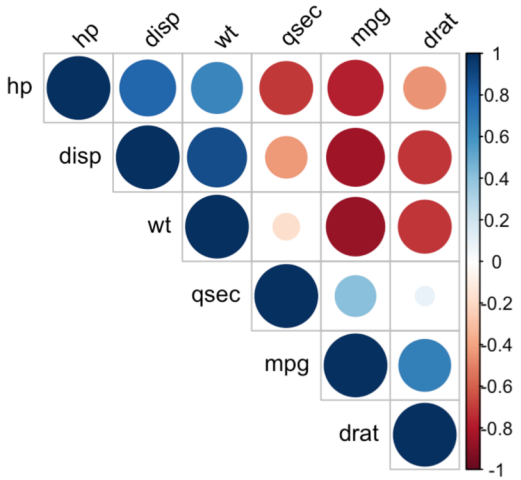
Heatmaps are a common way to simultaneously visualize multiple features of a dataset

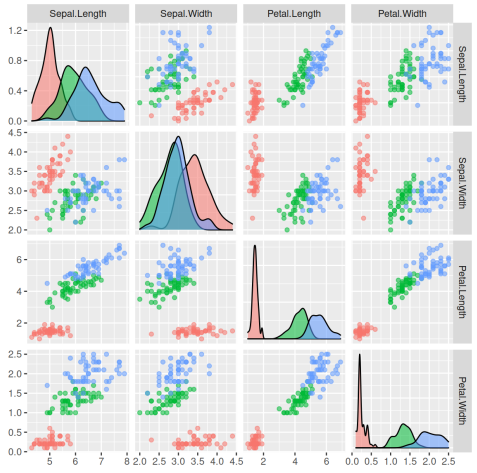


Dimensionality reduction

- “Dimensionality reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data”
  - Common approaches: principal component analysis (PCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), or non-negative matrix factorization (NMF) techniques, clustering by K-nearest neighbors algorithm
  - Used when dealing with large numbers of observations or variables when simpler approaches become unwieldy

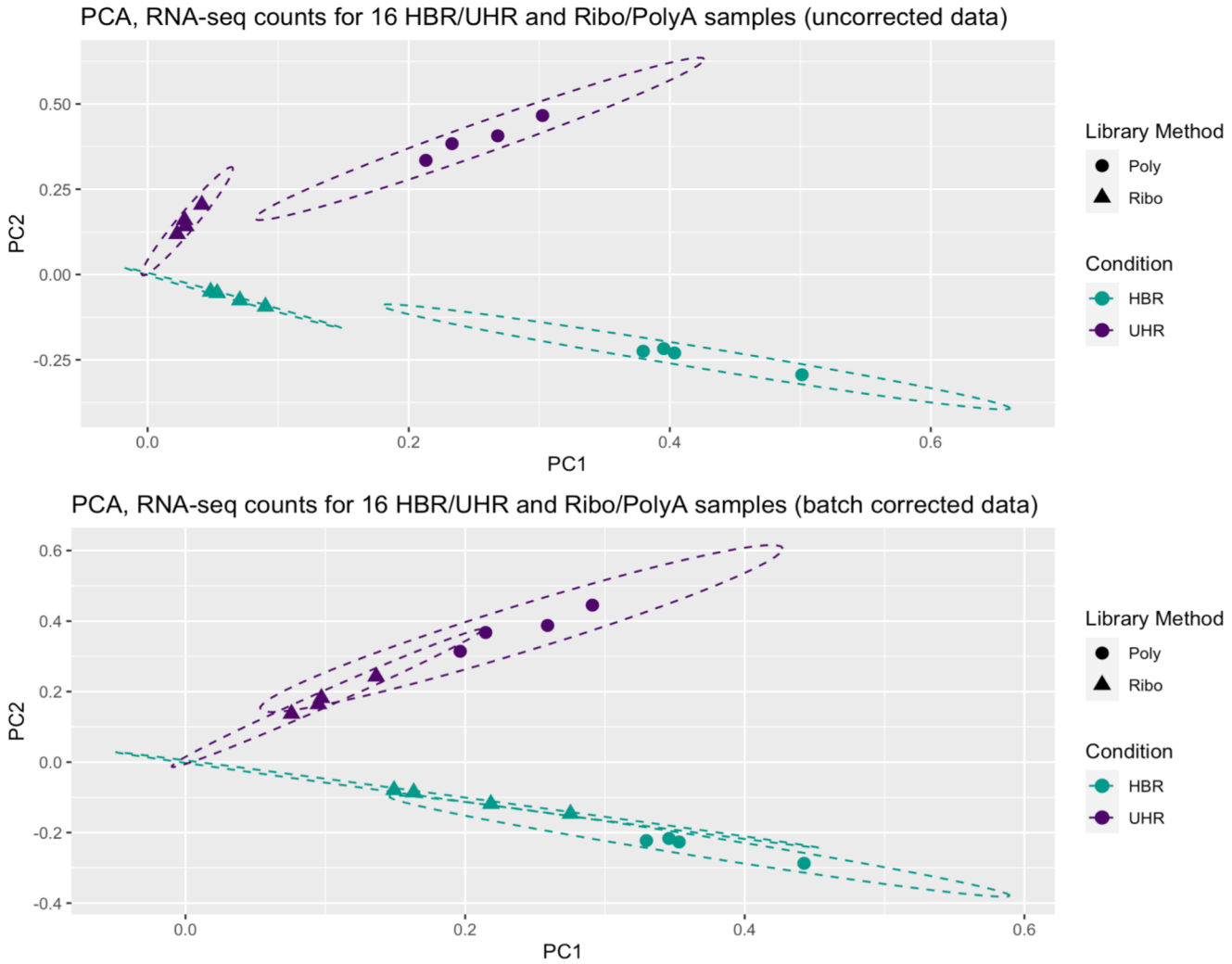






Dimensionality reduction (example 1 - RNA-seq batch effects)

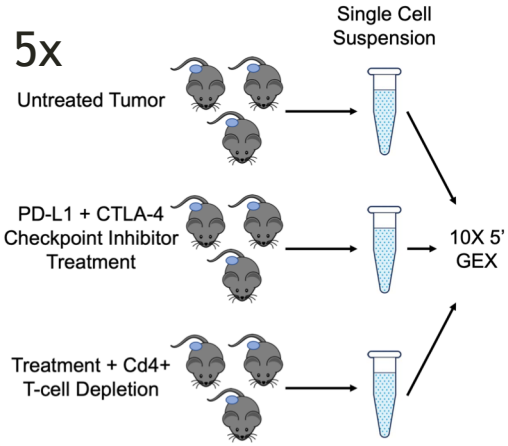
Principal component analysis (PCA) approach to dimensionality reduction applied to 16 bulk RNA-seq data sets. Used to visualize batch effects and batch correction

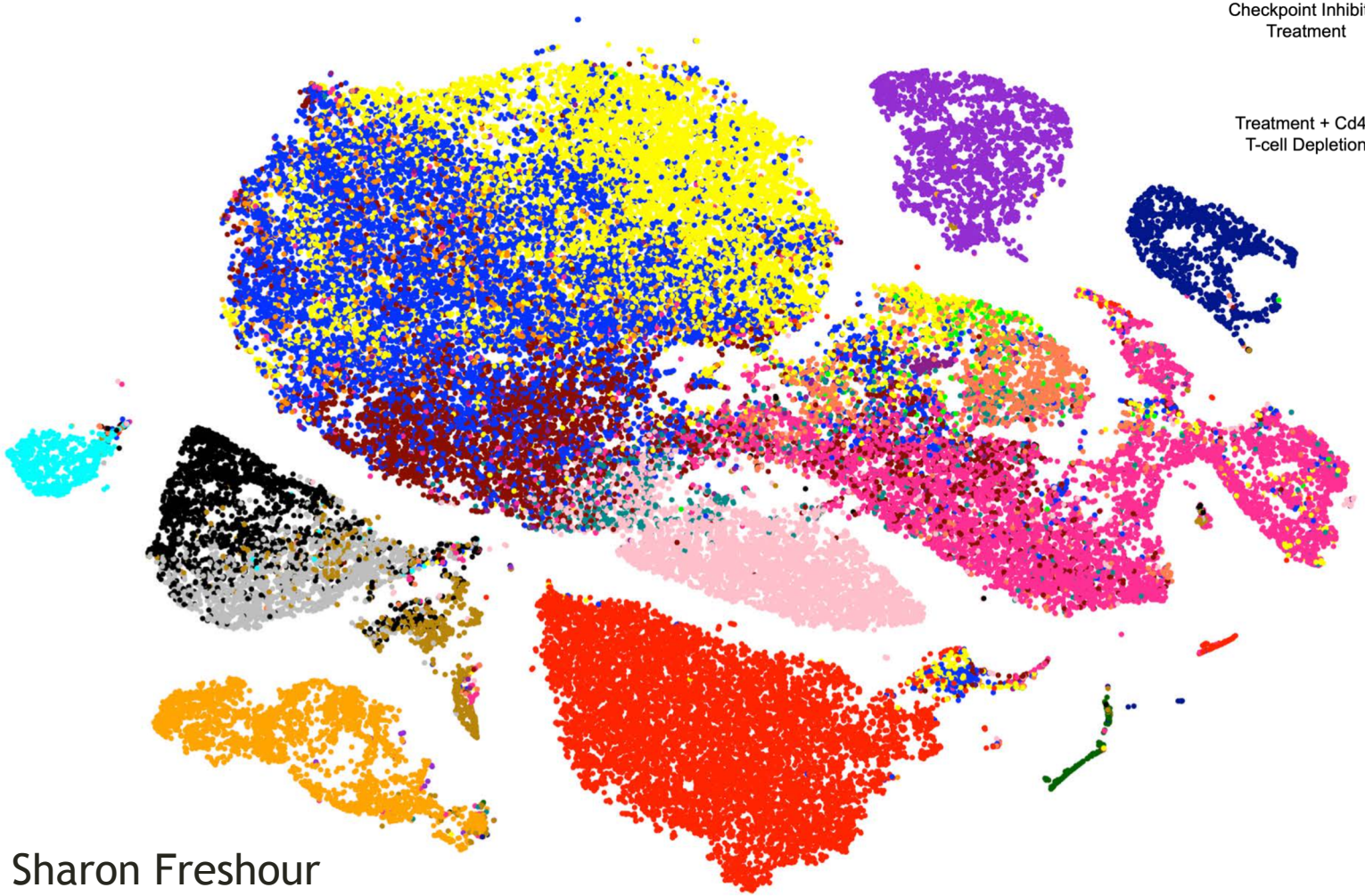


https://rnabio.org/module-03-expression/0003/05/01/Batch-Correction/

Dimensionality reduction (example 2 - single cell RNA)

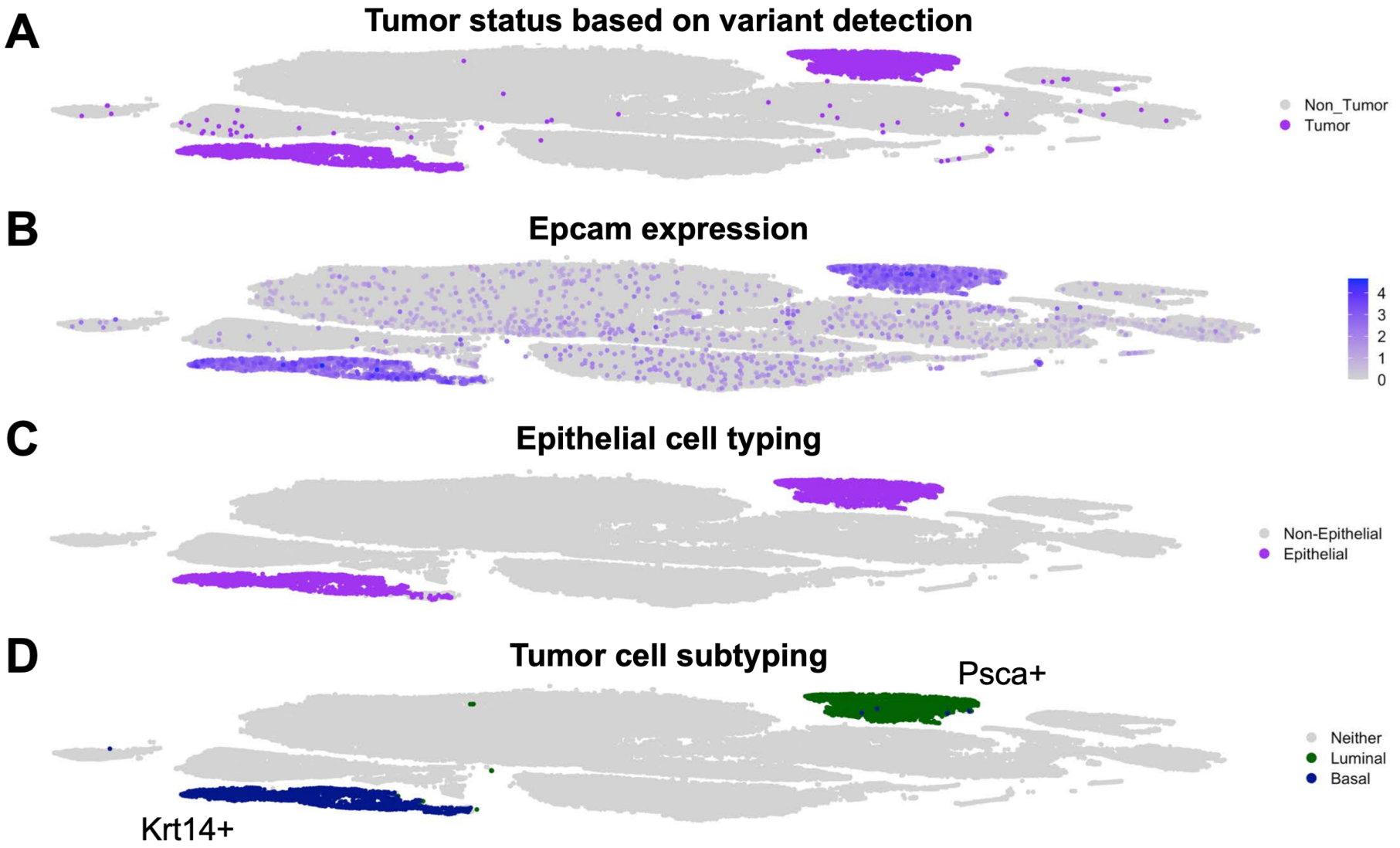
t-SNE approach to dimensionality reduction applied to visualize gene expression patterns from >60,000 cells





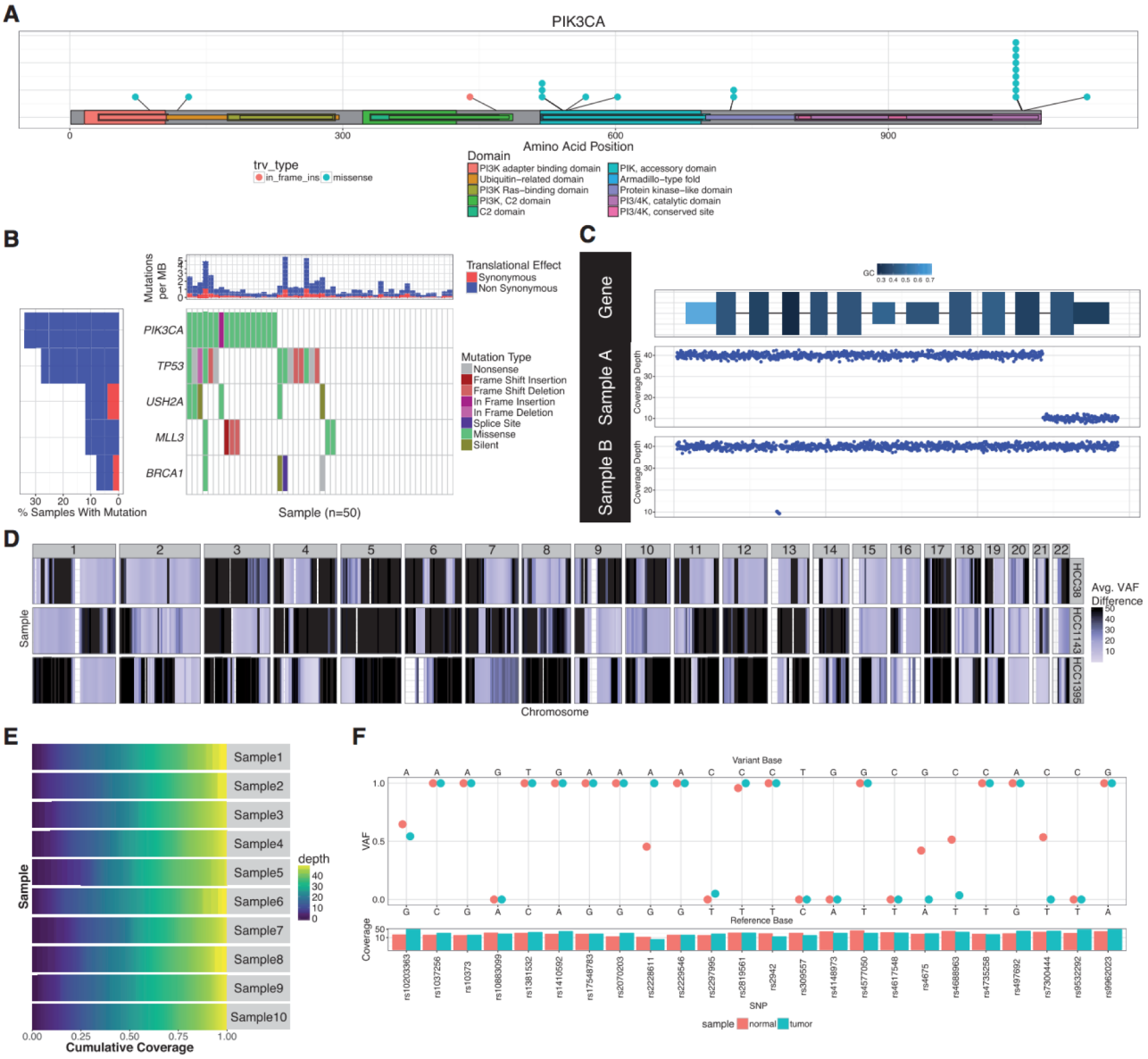
Sharon Freshour

Dimensionality reduction (example 2 - single cell RNA)



Sharon Freshour

GenVisR and genviz.org online course were created to help others make common genomic visualizations



Introduction to GenViz course site

www.genviz.org