Module 1B : Descriptive Statistics

Data Visualization

Agenda:

- Data types and their common visualizations:
    - Scatterplots
    - Mosaic and bar plots
    - Histograms
    - Box and Violin plots
    - Cumulative Frequency Distributions
- Interpretation of popular plots in genomics

Types of data:

Categorical Variable

- AKA Class variables or Nominal variables
- They do not have magnitude on a numerical scale
- **Nominal**
  - Lack inherent order
- **Ordinal**
  - Inherent order
- Ex: blood type, genotype, sex, state, survival (live or die), drug treatment (aspirin vs ibuprofen)

Quantitative Variables

- AKA Numerical variables
- Random Variable is a Quantitative variable
- **Continuous**
  - Ability to take any value ex.. Human weight, **age**
  - **They can be measured**
- **Discrete**
  - Spaces between possible values ex. Number of offspring, **age**
  - **They can be counted**

Data type determines plot type

- https://www.data-to-viz.com/ ←(and their code in Python and R)
- https://statisticsbyjim.com/graphs/
- https://piktochart.com/blog/types-of-graphs/
- https://www.sciencedirect.com/science/article/pii/S2666389920301896
- https://www.nature.com/articles/d41586-023-03393-9

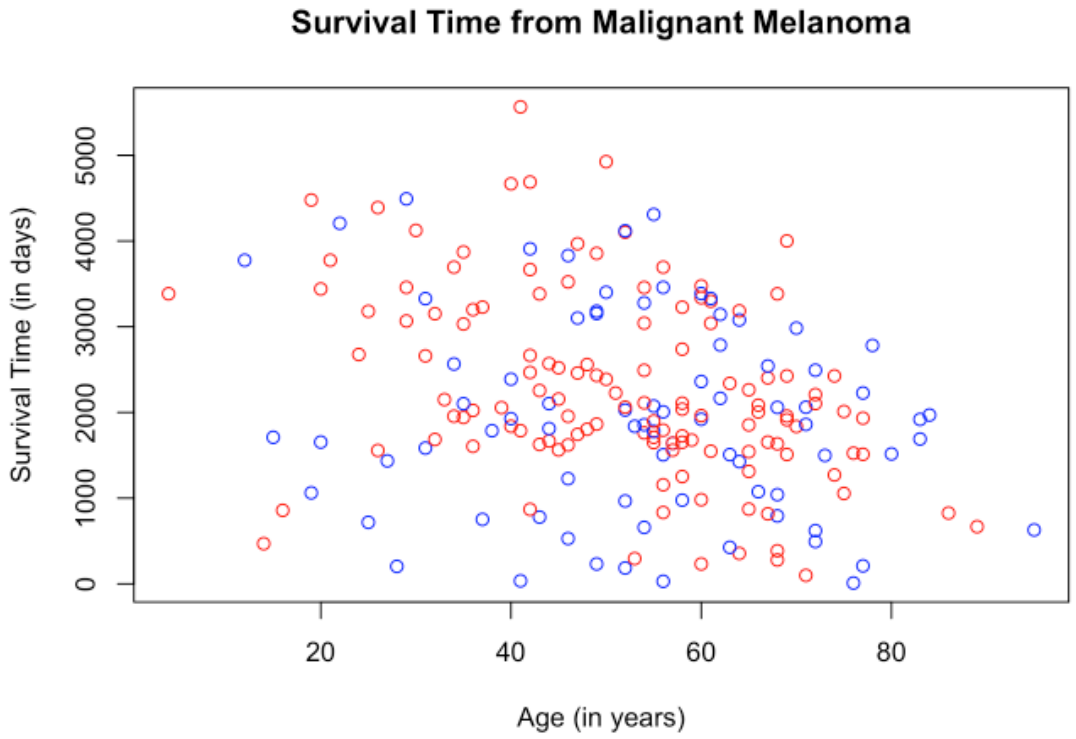
https://www.edwardtufte.com/tufte/
https://monachalabi.com/

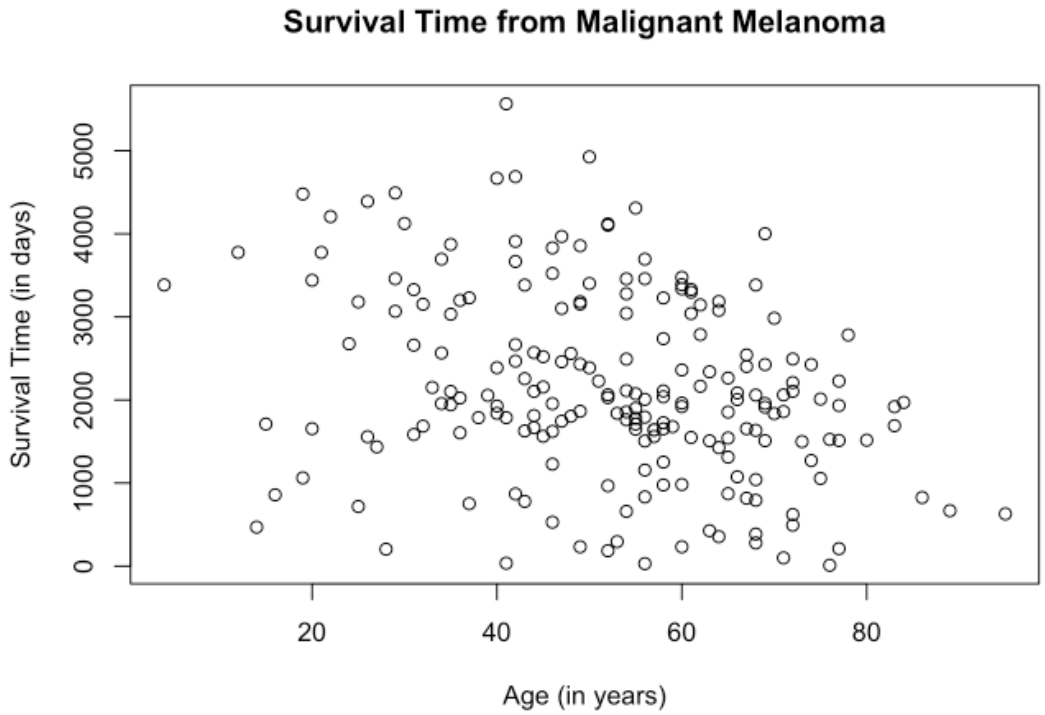
The plots we will examine:

- Scatterplots
- Histograms
- Mosaic plots, Bar plots
- Boxplots & Violin plots
- Cumulative Frequency Plots

Types of data	Graphical Method
Two numeric variables	• scatter
Two categorical variables	• Grouped Barplot • mosaic plot
one numerical variable, one categorical variable	• Violin plot / Box plot • Cumulative Freq Distr^ • multiple histograms

Scatterplot



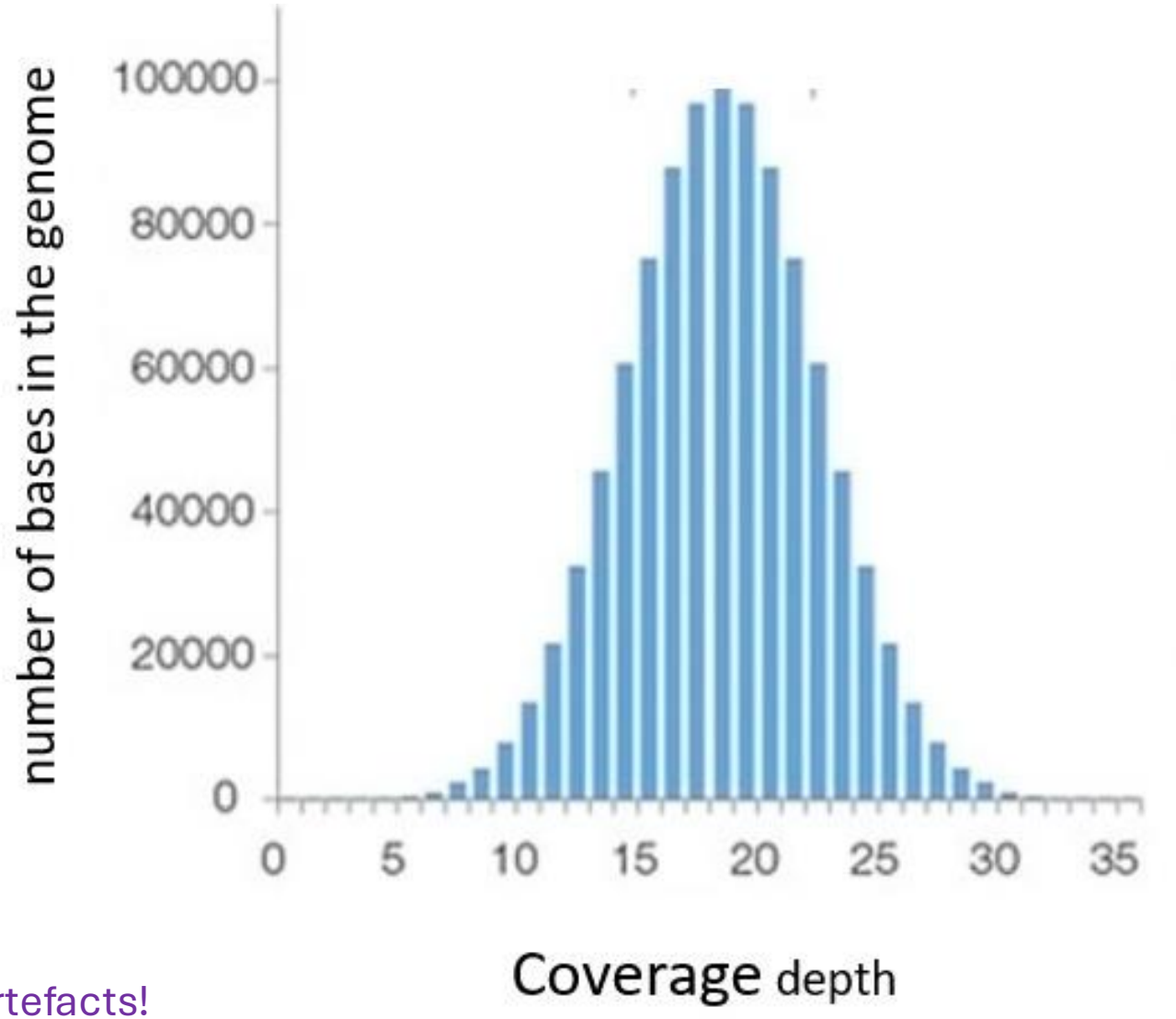


Free online textbook that gives r code!
https://bookdown.org/dli/rguide/scatterplots-and-best-fit-lines-two-sets.html

Hans Rosling ted talk (his website has data visualizations – scatterplots that move!- and datasets):
https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

Histogram

Coverage plot of complete genome

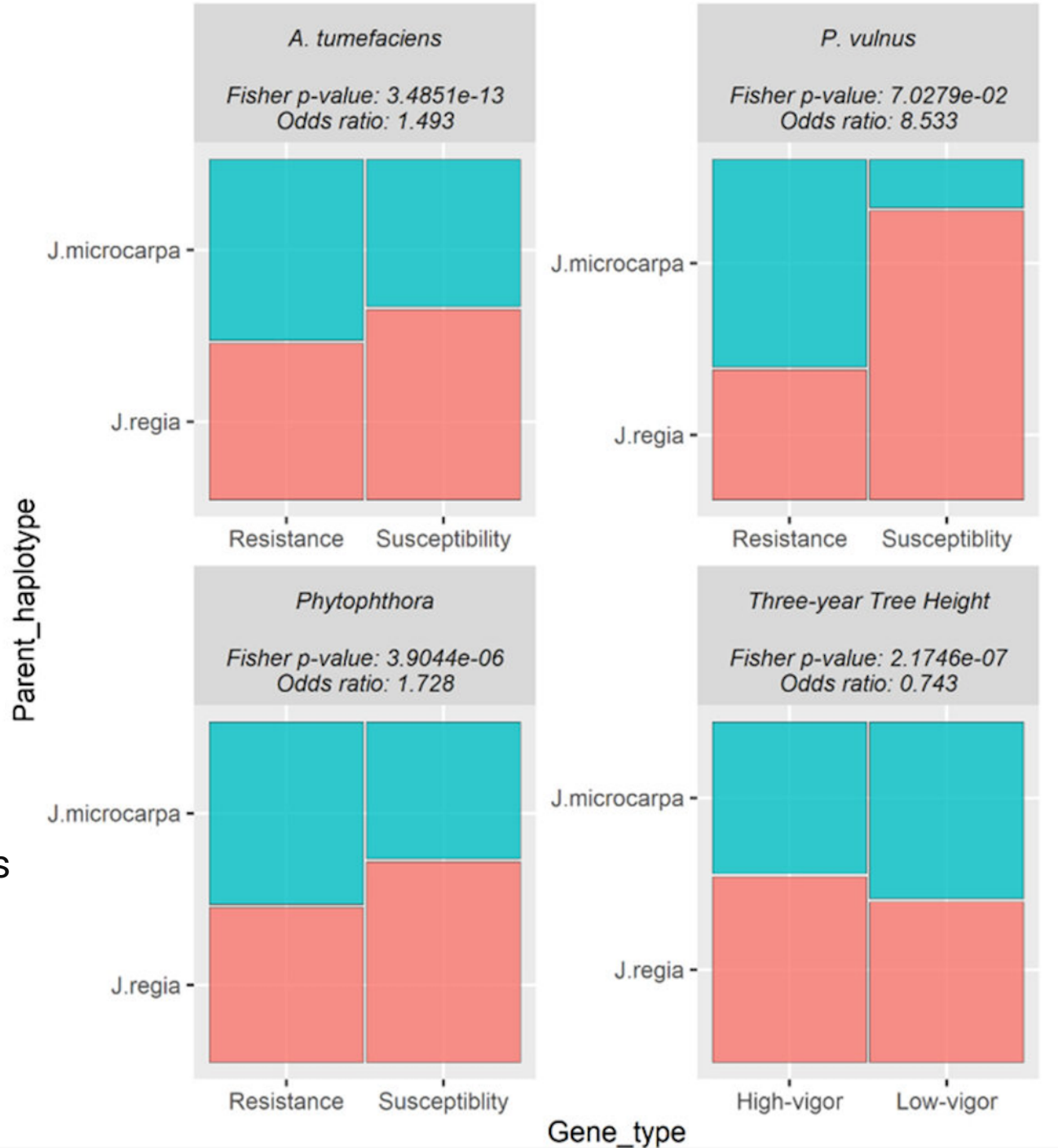


One warning about histograms:
Be careful about “bin” size; you can introduce artefacts!

https://www.biostars.org/p/9487269/

Mosaic Plot

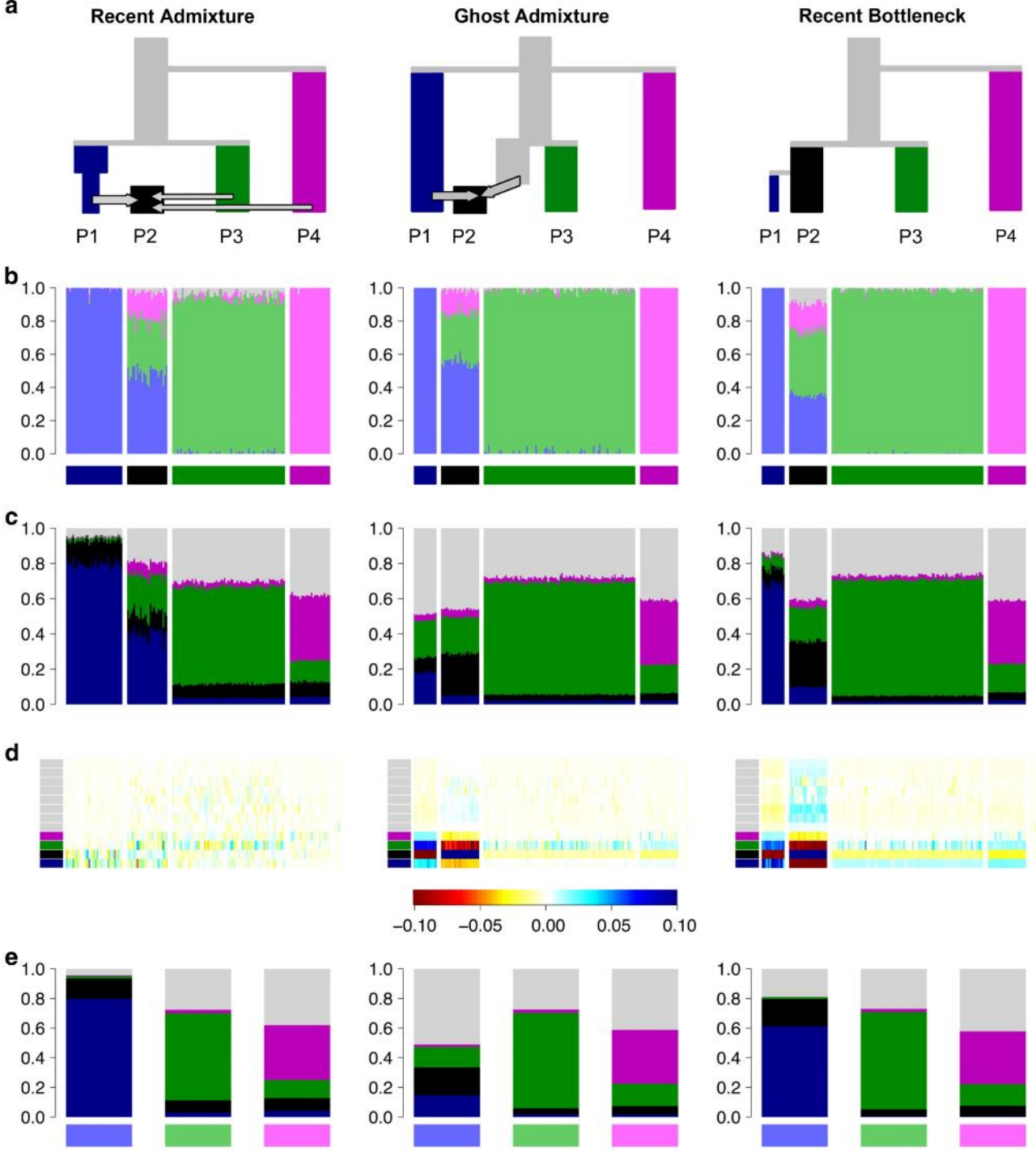
https://www.mdpi.com/1422-0067/25/2/931



Mosaic plot representing proportions of differentially expressed genes (DEGs) from each trait colored by the haplotype the genes mapped to. Each plot is labeled with the pathogen, Fisher's exact p-value, and Fisher's exact odds ratio. The odds ratio represents the ratio of the odds of the J. regia haplotype expressing a gene positively correlated to the trait compared to the odds of the J. microcarpa haplotype expressing a gene negatively correlated to the trait.

Bar Plot

https://www.nature.com/articles/s41467-018-05257-7

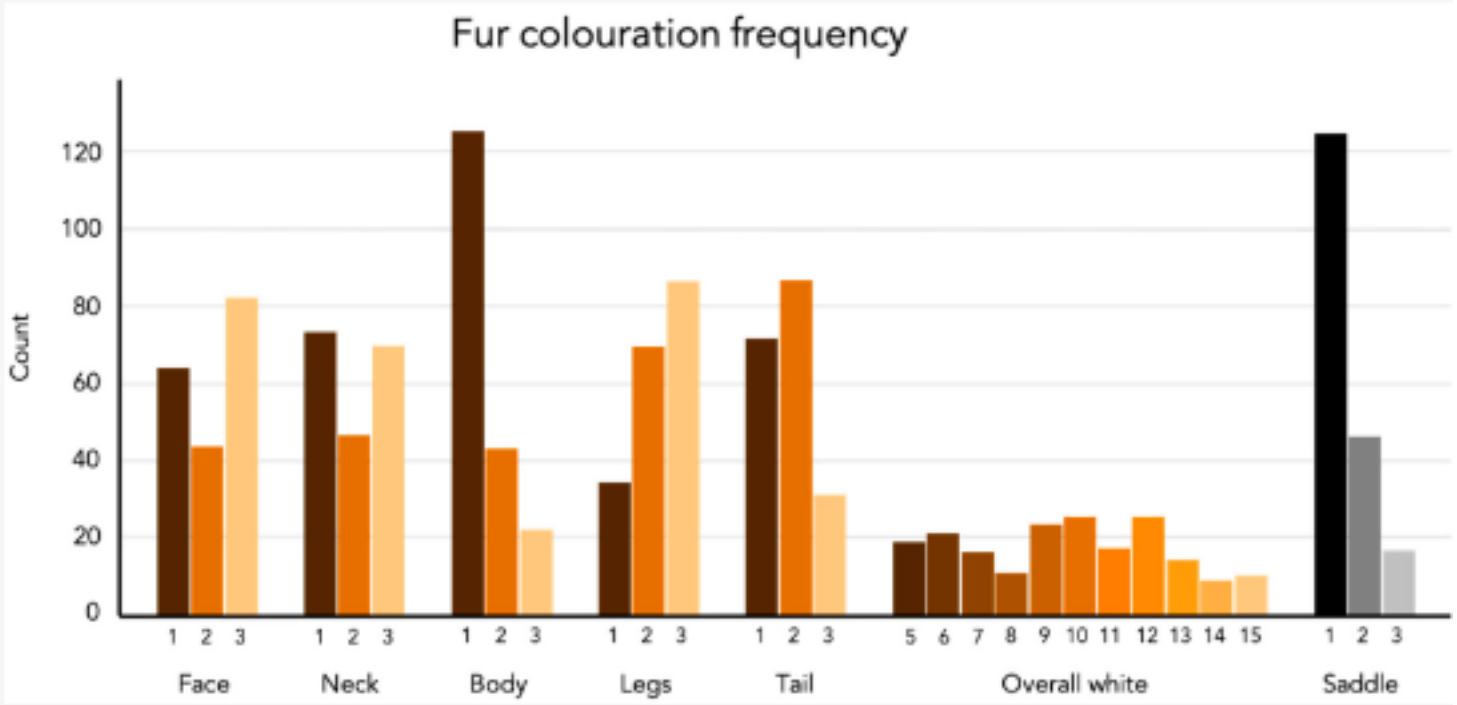


Three scenarios that give indistinguishable ADMIXTURE results. **a** Simplified schematic of each simulation scenario. **b** Inferred ADMIXTURE plots at $K = 11$. **c** CHROMOPAINTER inferred painting palettes. **d** Painting residuals after fitting optimal ancestral palettes using badMIXTURE, on the residual scale shown. **e** Ancestral palettes estimated by badMIXTURE. 13 populations in total were simulated, with grey populations all being outgroups to those shown in colour

Bar Plot

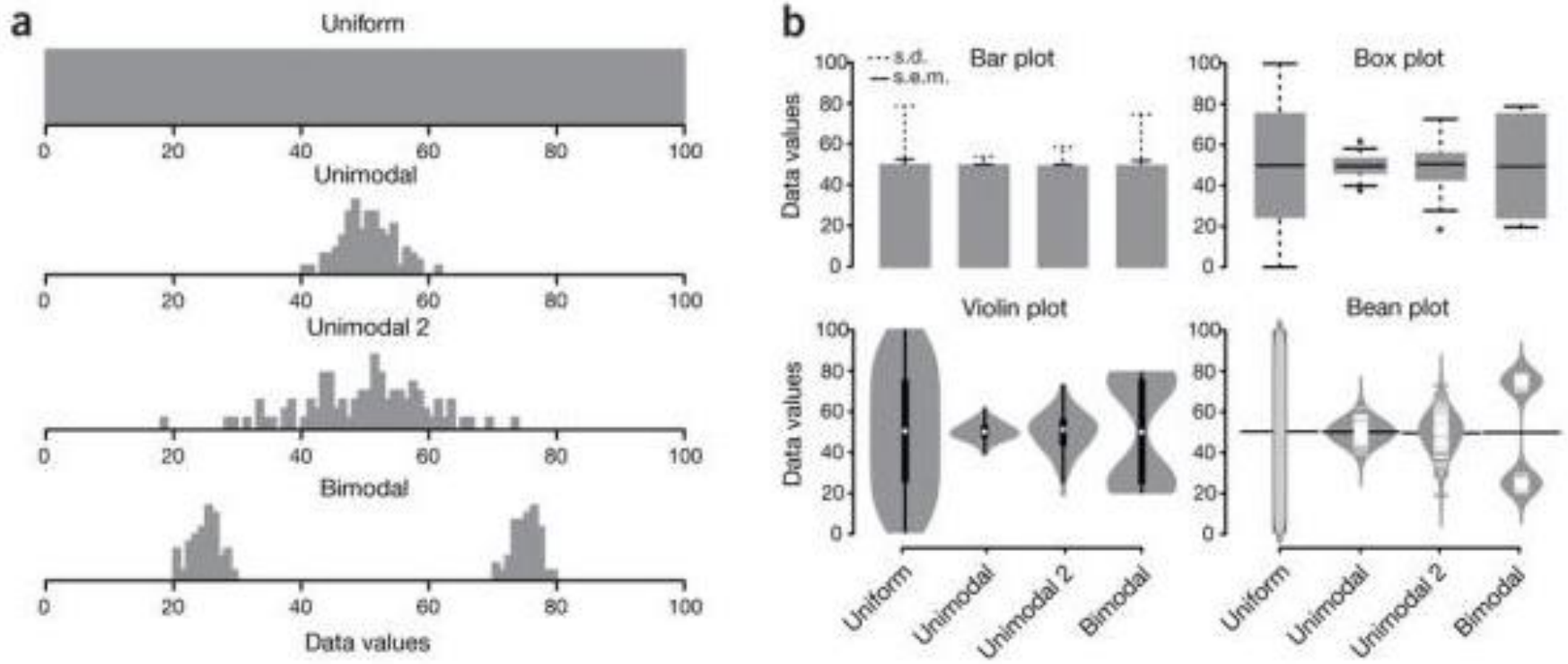
https://www.mdpi.com/2073-4425/12/2/316

Figure 1. The frequency (count) of individuals for each phenotype scoring. The total number of individuals is 190 with Table 187 due to the exclusion of three dogs that did not express the saddle.



The frequency (count) of individuals for each phenotype scoring. The total number of individuals is 190 with Table 187 due to the exclusion of three dogs that did not express the saddle

Boxplots & Violin plots



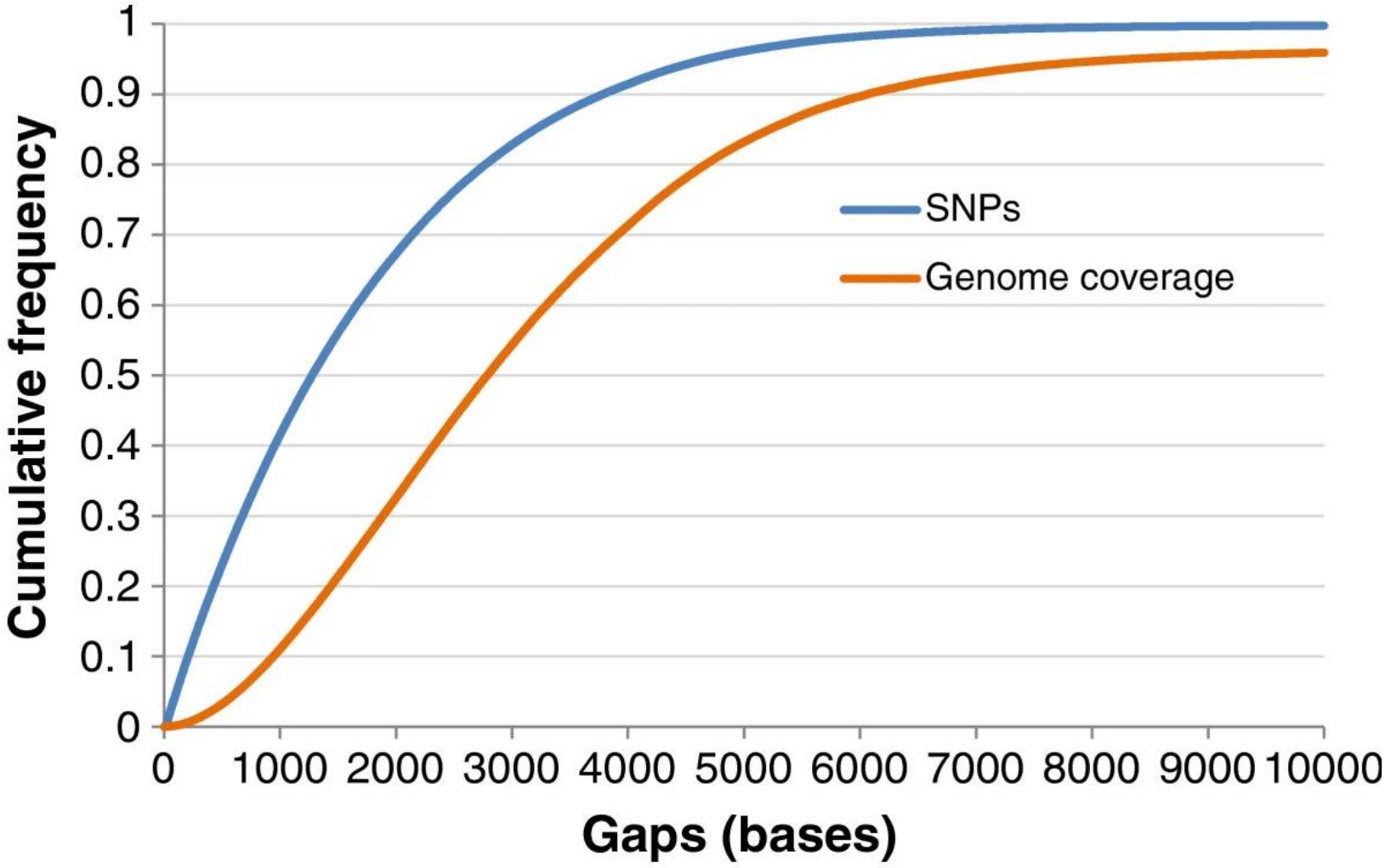
Data visualization with box plots(a)

Hypothetical sample data sets of 100 data points each that are uniform, unimodal with one of two different variances or bimodal. Simple bar plot representations and statistical parameters may obscure such different data distributions.

(b) Comparison of data visualization methods. Bar plots typically represent only the mean and s.d. or s.e.m. Box plots visualize the five-number summary of a data set (minimum, lower quartile, median, upper quartile and maximum). Violin and bean plots represent the actual distribution of the individual data sets.

https://pubmed.ncbi.nlm.nih.gov/24481215/

Cumulative Frequency Distribution



https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-59

Cumulative frequency distributions of SNPs and genome coverage as functions of inter-marker spacing in the panel. Inter-marker spacing included distances between consecutive SNPs and the distances from chromosome ends to the nearest SNP in 600 K panel.

Question: The table below shows the frequency of a specific single nucleotide polymorphism (SNP) across four different populations:

Population	Frequency
A	15
B	25
C	45
D	60

Based on the data provided, which of the following visualizations would best represent the differences in SNP frequency across populations?

A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot

Question: The following data represents the total number of mutations detected in a genetic sample as a function of increasing sequencing depth:

Sequencing Depth	Cumulative Mutations Detected
10X	15
20X	40
30X	70
40X	100
50X	120

Which of the following visualizations would best represent the accumulation of detected mutations as sequencing depth increases?

A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot

Question: The table below shows the observed genotype frequencies of a SNP in two populations:

Population	Genotype AA	Genotype AB	Genotype BB
A	40%	35%	25%
B	30%	50%	20%

Which of the following visualizations would best represent the proportional relationship of genotypes within and between populations?

A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot

Mouse_ID	Strain	Gene_Expression (GFP)	Mutation_Type	Age (weeks)	Sex
M001	C57BL/6J	12.4	None	10	Male
M002	BALB/c	8.1	SNP	12	Female
M003	C57BL/6J	15.6	InDel	14	Male
M004	DBA/2J	7.3	None	9	Female
M005	BALB/c	9.9	SNP	11	Male
M006	C57BL/6J	14.7	InDel	13	Female
M007	DBA/2J	6.8	None	8	Male

Question A: You want to compare the **distribution of gene expression levels** across different **mouse strains**. What is the most appropriate plot?

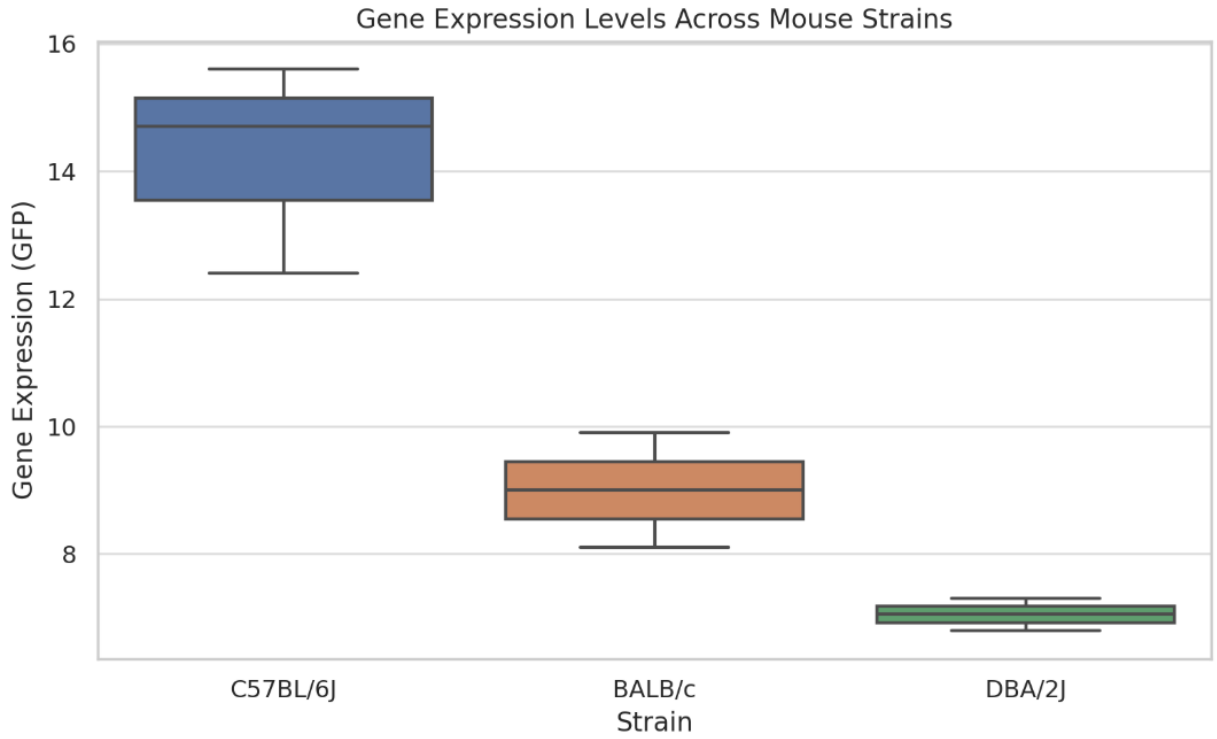
A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot

Mouse_ID	Strain	Gene_Expression (GFP)	Mutation_Type	Age (weeks)	Sex
M001	C57BL/6J	12.4	None	10	Male
M002	BALB/c	8.1	SNP	12	Female
M003	C57BL/6J	15.6	InDel	14	Male
M004	DBA/2J	7.3	None	9	Female
M005	BALB/c	9.9	SNP	11	Male
M006	C57BL/6J	14.7	InDel	13	Female
M007	DBA/2J	6.8	None	8	Male

Question A: You want to compare the **distribution of gene expression levels** across different **mouse strains**.

What is the most appropriate plot?

A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot



Mouse_ID	Strain	Gene_Expression (GFP)	Mutation_Type	Age (weeks)	Sex
M001	C57BL/6J	12.4	None	10	Male
M002	BALB/c	8.1	SNP	12	Female
M003	C57BL/6J	15.6	InDel	14	Male
M004	DBA/2J	7.3	None	9	Female
M005	BALB/c	9.9	SNP	11	Male
M006	C57BL/6J	14.7	InDel	13	Female
M007	DBA/2J	6.8	None	8	Male

Question B: You want to visualize the **relationship between Mutation_Type and Sex** across your mice. What plot is most appropriate?

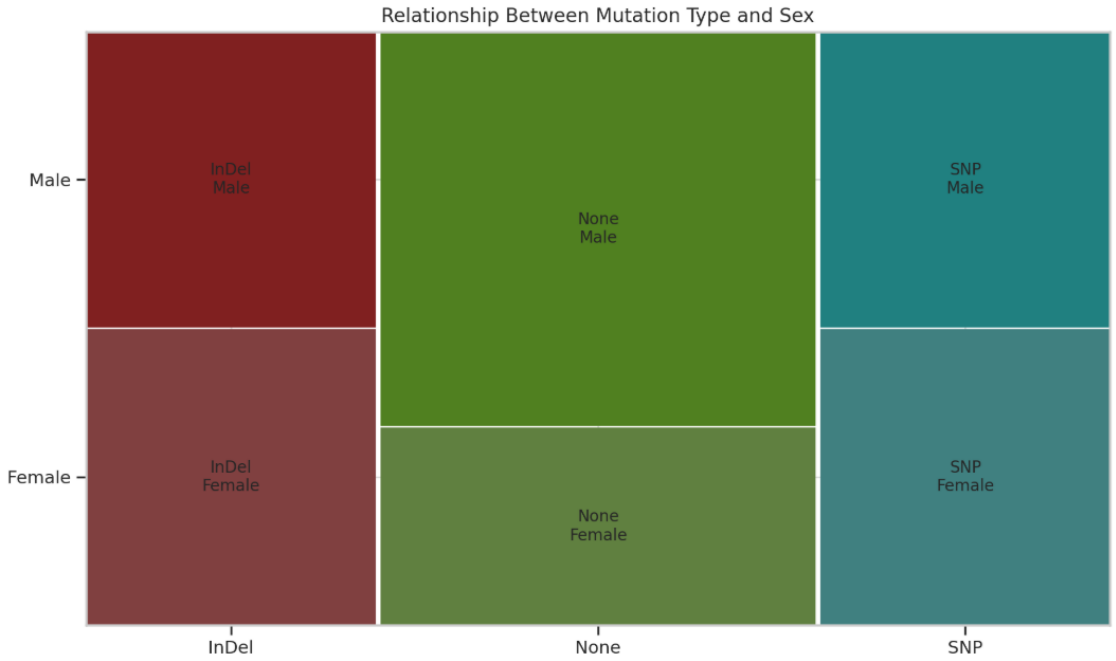
A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot

Mouse_ID	Strain	Gene_Expression (GFP)	Mutation_Type	Age (weeks)	Sex
M001	C57BL/6J	12.4	None	10	Male
M002	BALB/c	8.1	SNP	12	Female
M003	C57BL/6J	15.6	InDel	14	Male
M004	DBA/2J	7.3	None	9	Female
M005	BALB/c	9.9	SNP	11	Male
M006	C57BL/6J	14.7	InDel	13	Female
M007	DBA/2J	6.8	None	8	Male

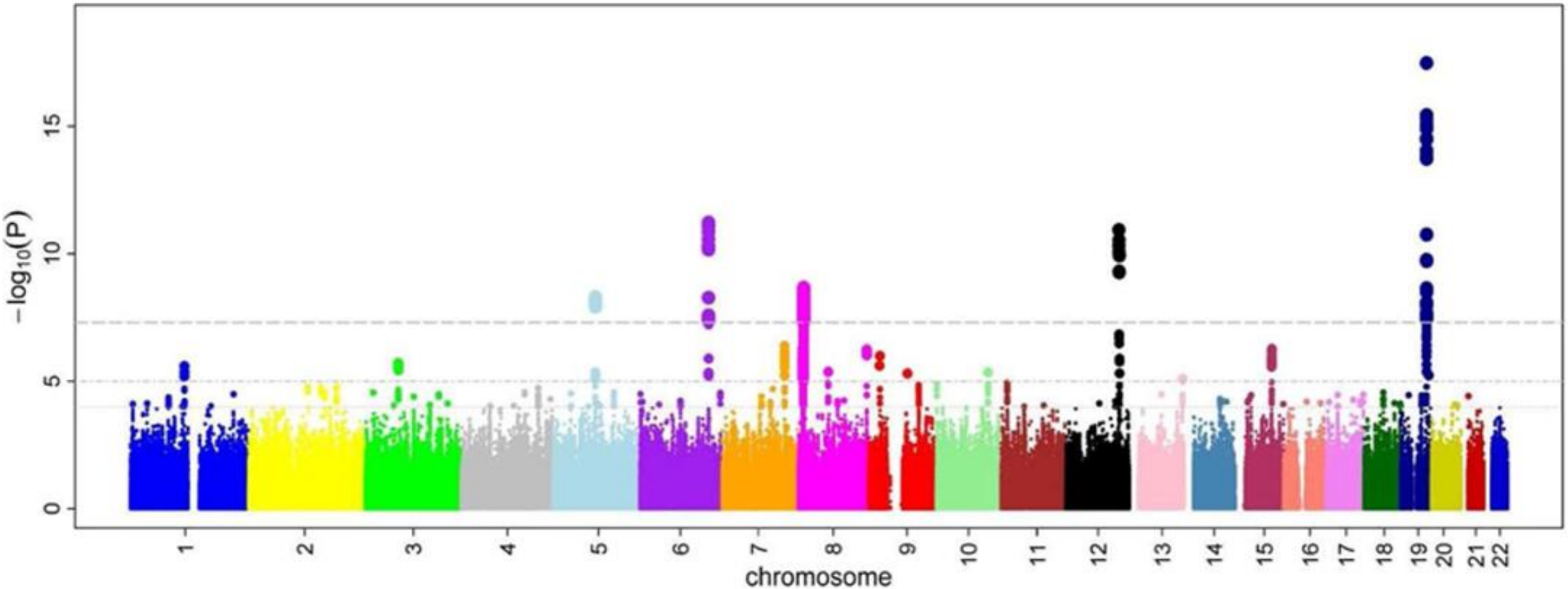
Question B: You want to visualize the **relationship between Mutation_Type and Sex** across your mice.

What plot is most appropriate?

A. Scatterplot
B. Cumulative frequency plot
C. Bar plot
D. Mosaic plot
E. Boxplot



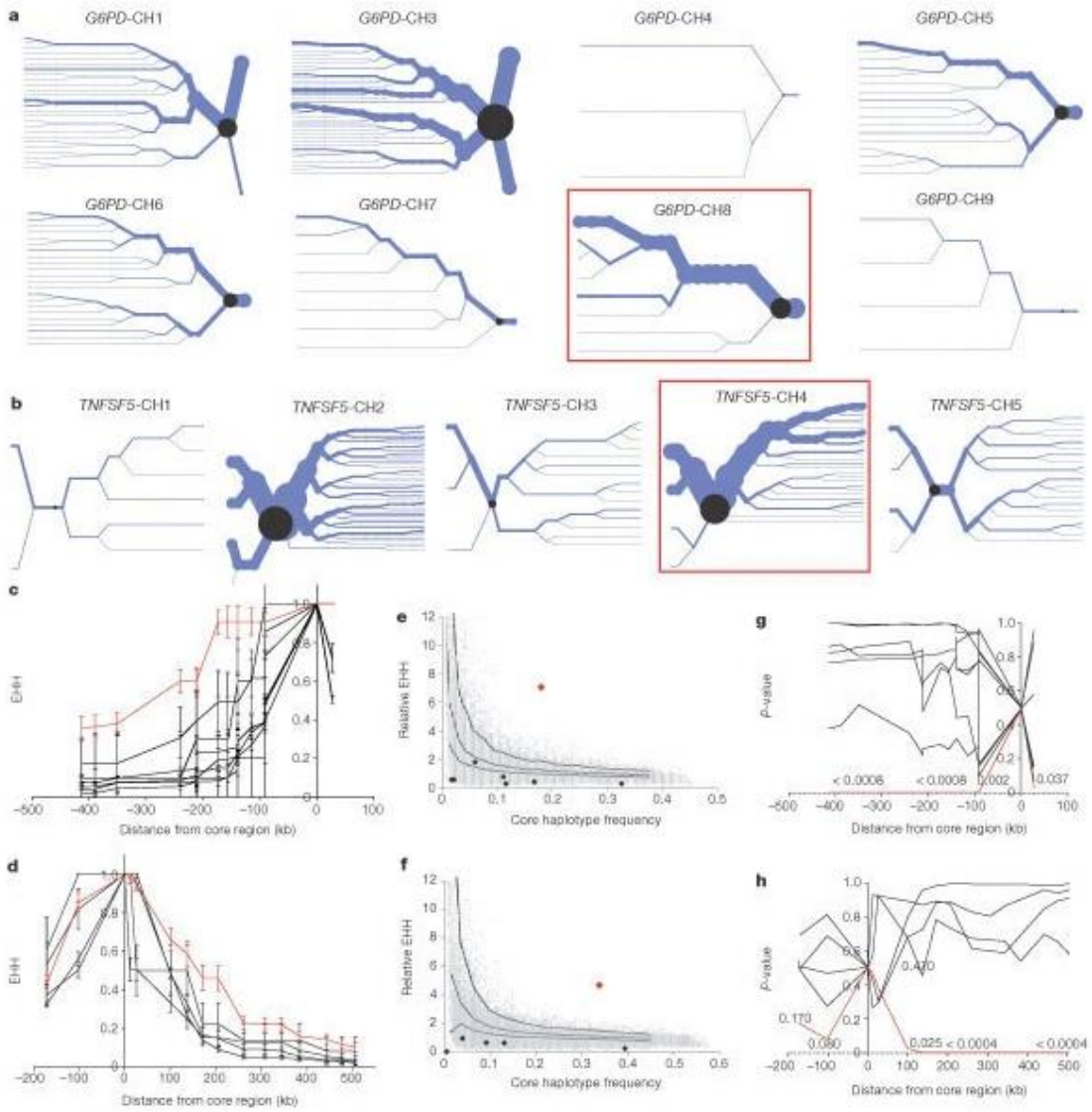
Interpret the following plots



https://en.wikipedia.org/wiki/Manhattan_plot

Detecting recent positive selection in the human genome from haplotype structure

a, **b**, Haplotype bifurcation diagrams (see Methods) for each core haplotype at *G6PD* (**a**) and *TNFSF5* (**b**) in pooled African populations demonstrate that *G6PD*-CH8 and *TNFSF5*-CH4 (boxed or labelled in red) have long-range homozygosity that is unusual given their frequency. **c**, **d**, The EHH at varying distances from the core region on each core haplotype at *G6PD* (**c**) and *TNFSF5* (**d**) demonstrates that *G6PD*-CH8 and *TNFSF5*-CH4 have persistent, high EHH values. **e**, **f**, At the most distant SNP from *G6PD* (**e**) and *TNFSF5* (**f**) core regions, the relative EHH plotted against the core haplotype frequency is presented and compared with the distribution of simulated core haplotypes (on the basis of simulation of 5,000 data sets; represented by grey dots and given with 95th, 75th and 50th percentiles). The observed non-selected core haplotypes in our data are represented by black diamonds. **g**, **h**, We calculated the statistical significance of the departure of the observed data from the simulated distribution at each distance from the core. *G6PD*-CH8 (**g**) and *TNFSF5*-CH4 (**h**) demonstrate increasing deviation from a model of neutral drift at further distances from the core region in both directions

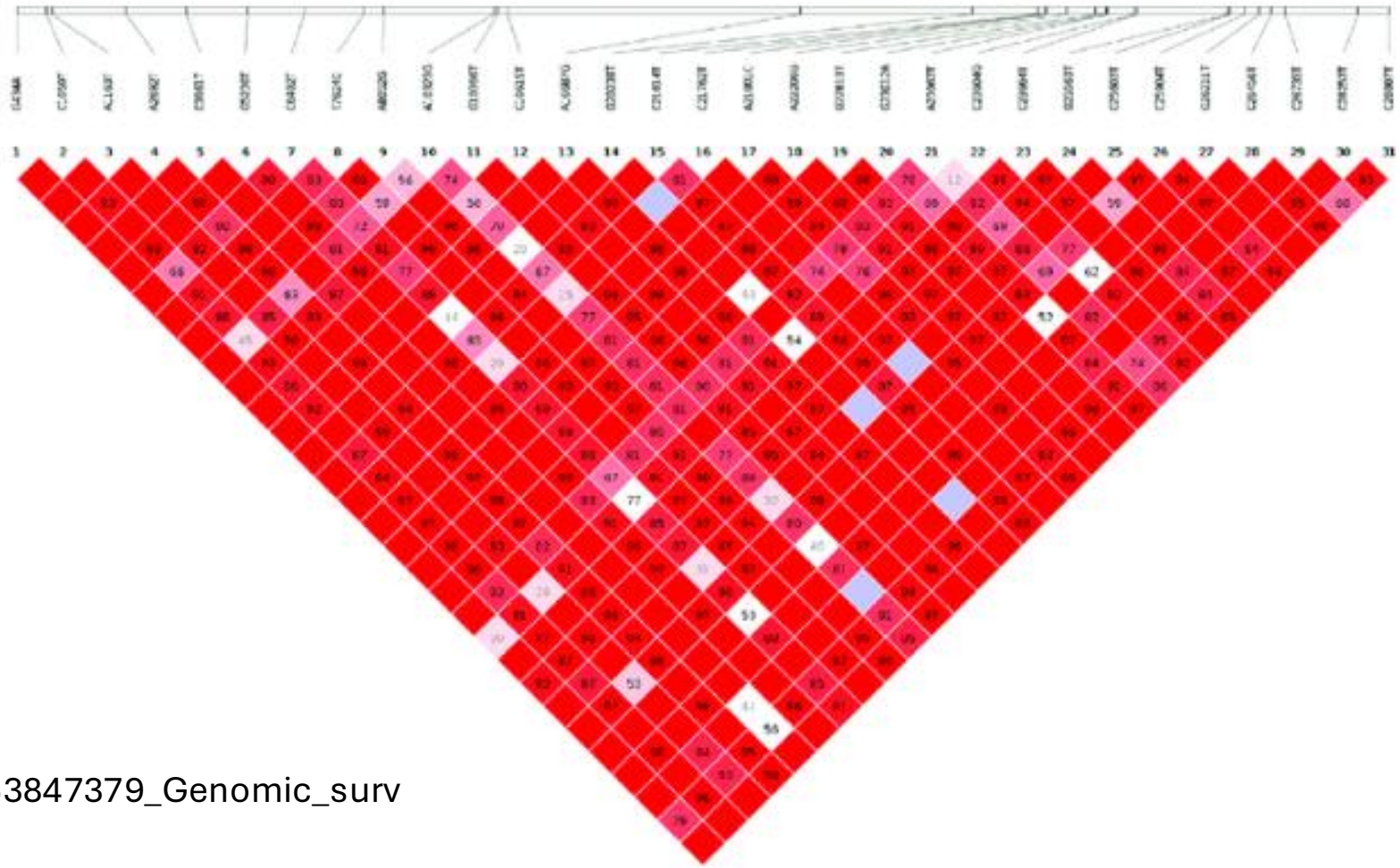


https://software.broadinstitute.org/mpg/sweep/

https://www.nature.com/articles/nature01140

(Sabeti et al, 2002)

Textile Plot



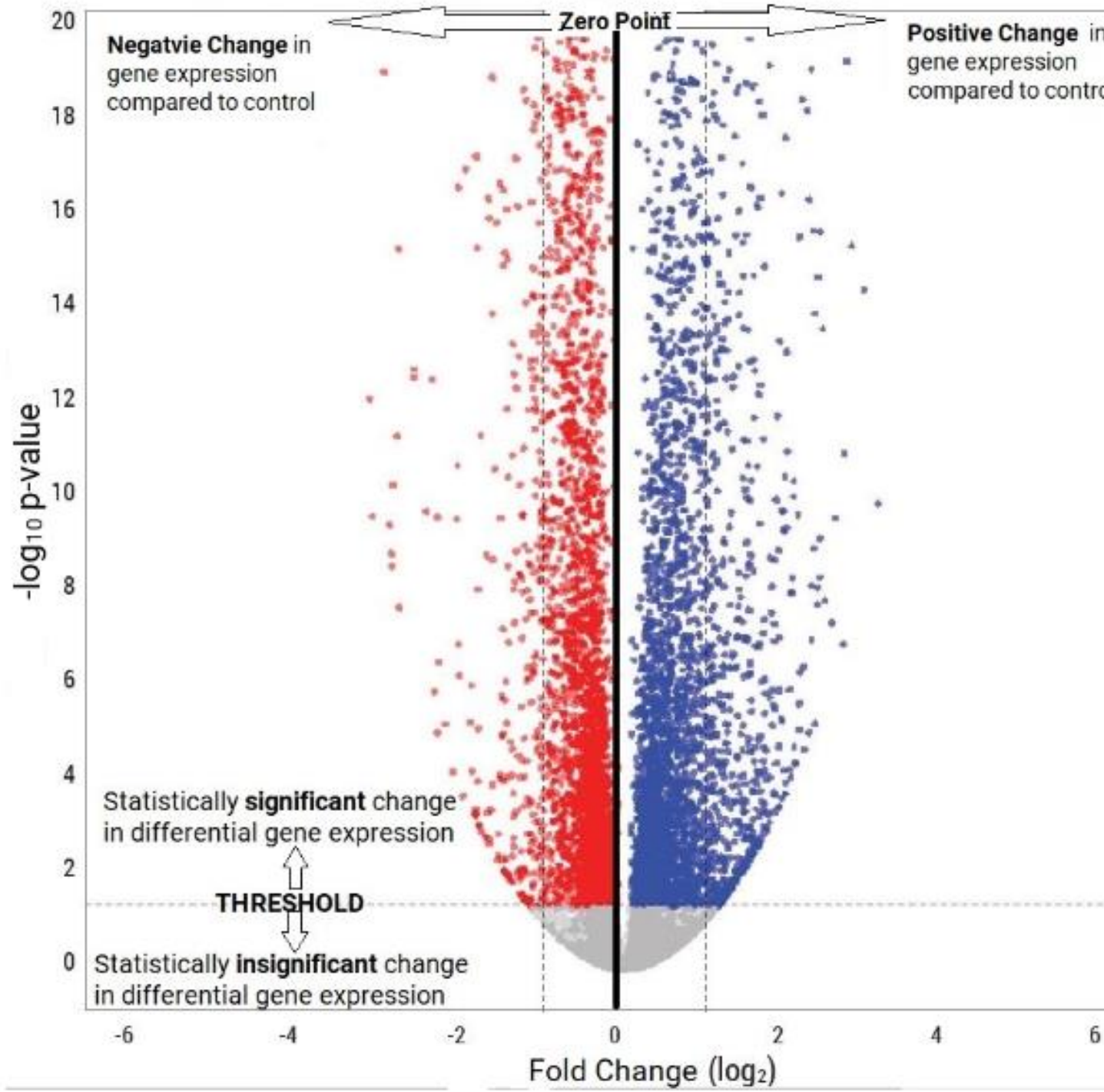
Originally described here:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010207

https://www.researchgate.net/publication/363847379_Genomic_surveillance_unfolds_the_SARS-CoV-2_transmission_and_divergence_dynamics_in_Bangladesh

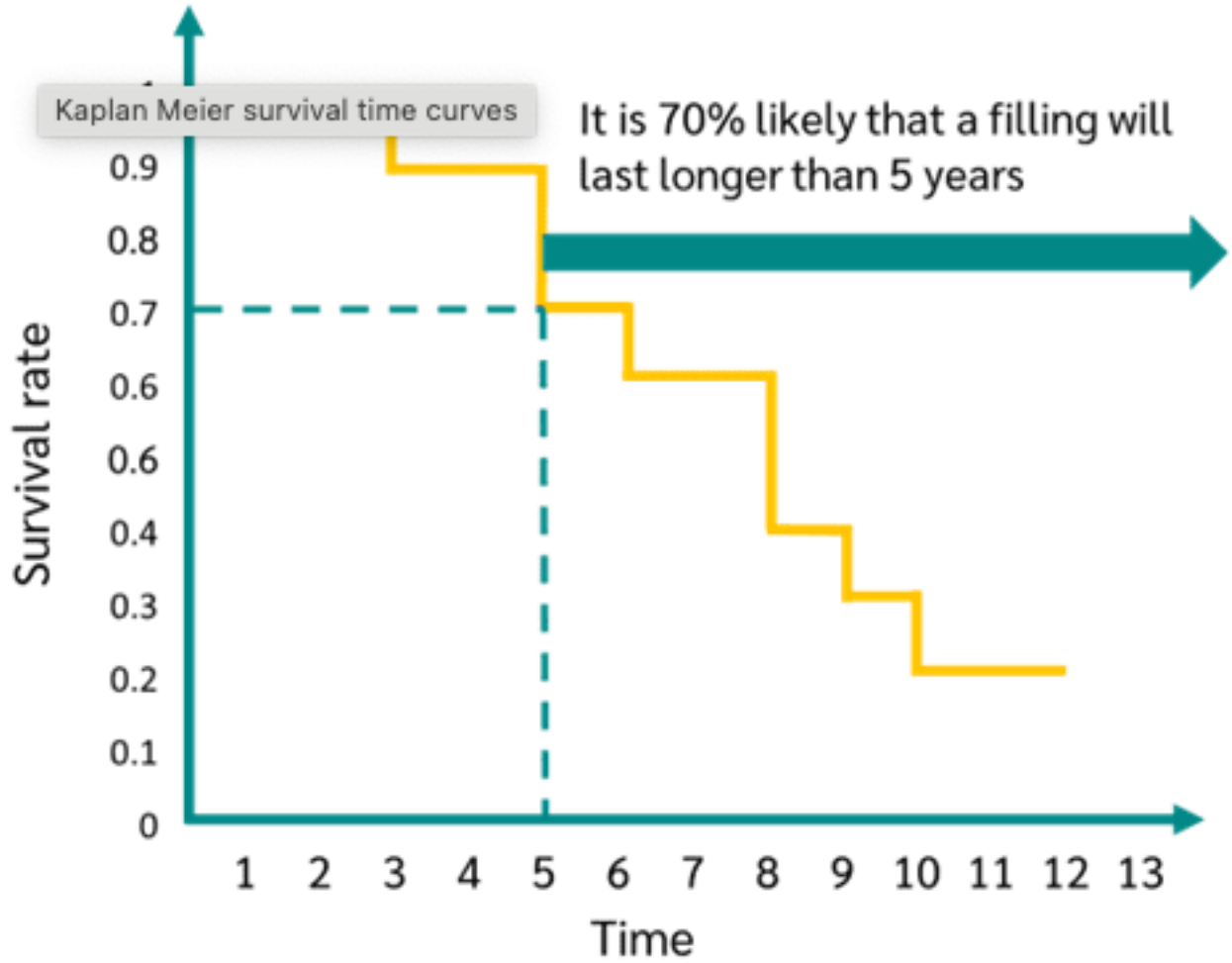
Linkage disequilibrium plot. The LD plot is generated considering the most prevalent SNPs. The number at the top denotes the SNP position, and squares are colored by standard (D'/LOD). The brighter red color indicates a higher D' value and vice versa. The number in square is r 2 value.

Volcano plots like the one shown above are useful when there are many (thousands or even millions) of observations with a wide range of differences, both positive and negative. It exhibits a densely populated, symmetrical “V” shape. When the number of observations is reduced or the variation in response is not so evenly distributed, the volcano plot might appear as shown below.

https://www.htgmolecular.com/blog/2022-08-25/understanding-volcano-plots



https://datatab.net/tutorial/kaplan-meier-curve



Summary

1. The appropriate visualization will depend on the type of variable(s) you are graphing

# variables	Variable Type	Recommended Plots	Use Case
1 (univariate)	Categorical	Bar Chart~~, Pie Chart~~	Comparing category frequencies
	Numerical	Histogram, Boxplot, Density Plot	Understanding distributions
2 (Bivariate)	Categorical & Categorical	Grouped Bar Chart, Mosaic Plot	Comparing proportions of two groups
	Numerical & Categorical	Boxplot, Violin Plot, Strip Plot	Comparing distributions across categories
	Numerical & Numerical	Scatter Plot, Line Plot, Hexbin Plot	Examining relationships or trends
3+ (Multivariate)	Multiple Categorical	Stacked Bar Chart	Analyzing categorical interactions
	Multiple Numerical	Scatterplot Matrix	Comparing multiple numeric relationships
	Mixed	Faceted Plots, Heatmap, Bubble Chart	Visualizing mixed data relationships

2. Everything else is (mostly) artistry and **being clear** in what you are revealing to your audience (See: Edward Tufte for “rules”)