Module 1 : Descriptive Statistics

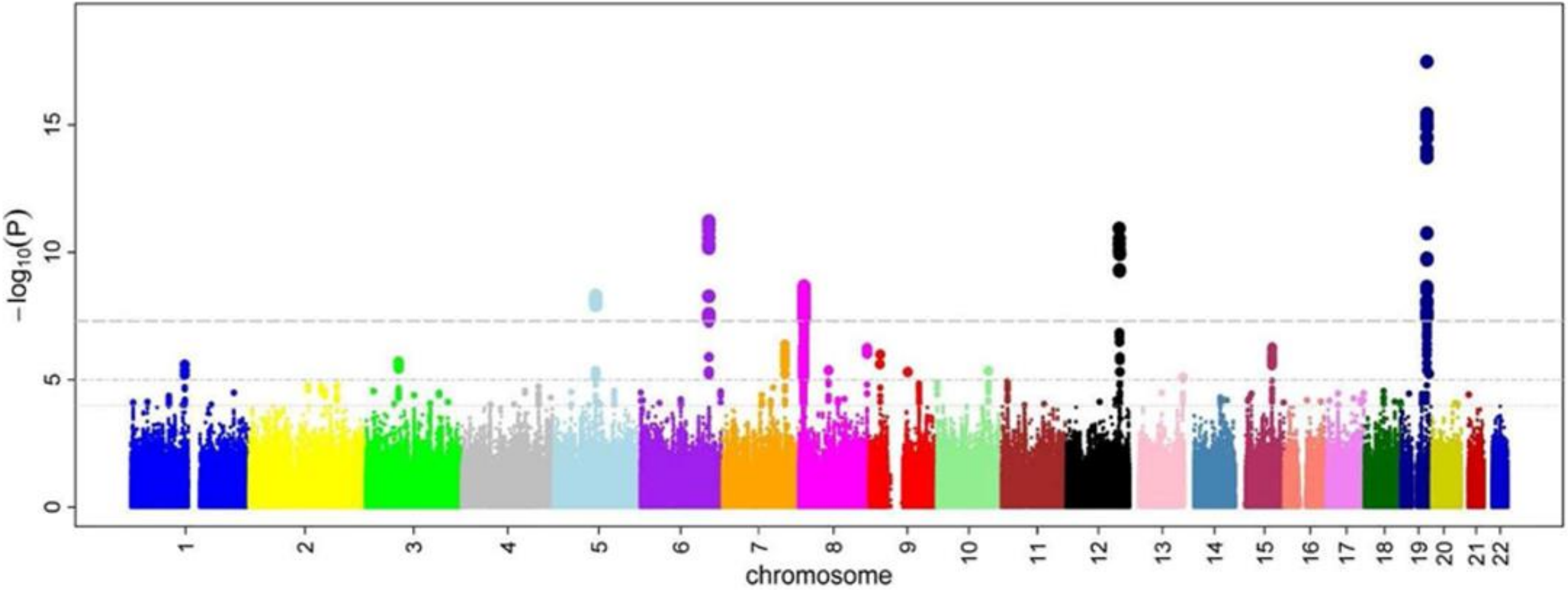
Data Visualization

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**

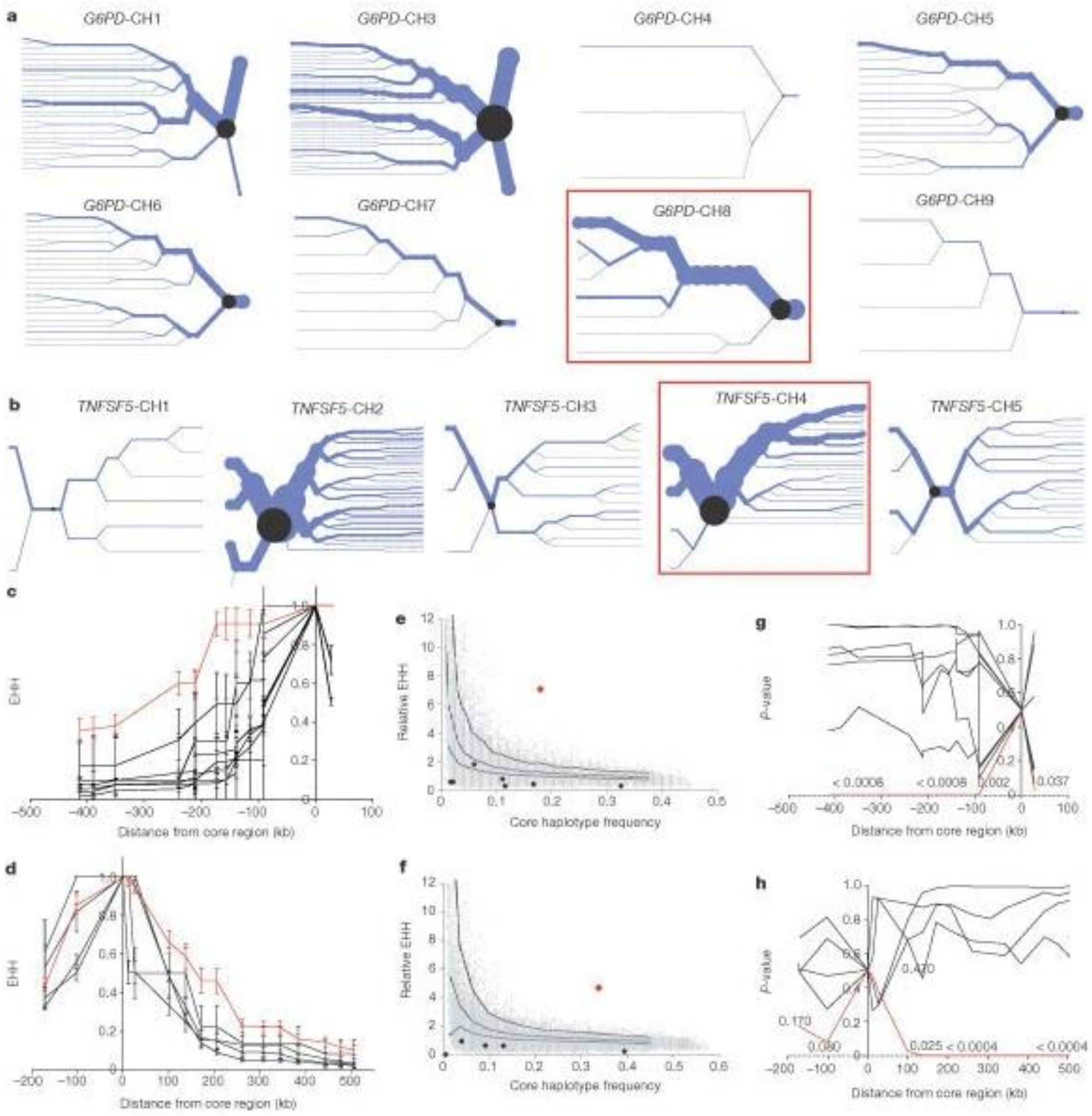
Interpret the following plots



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



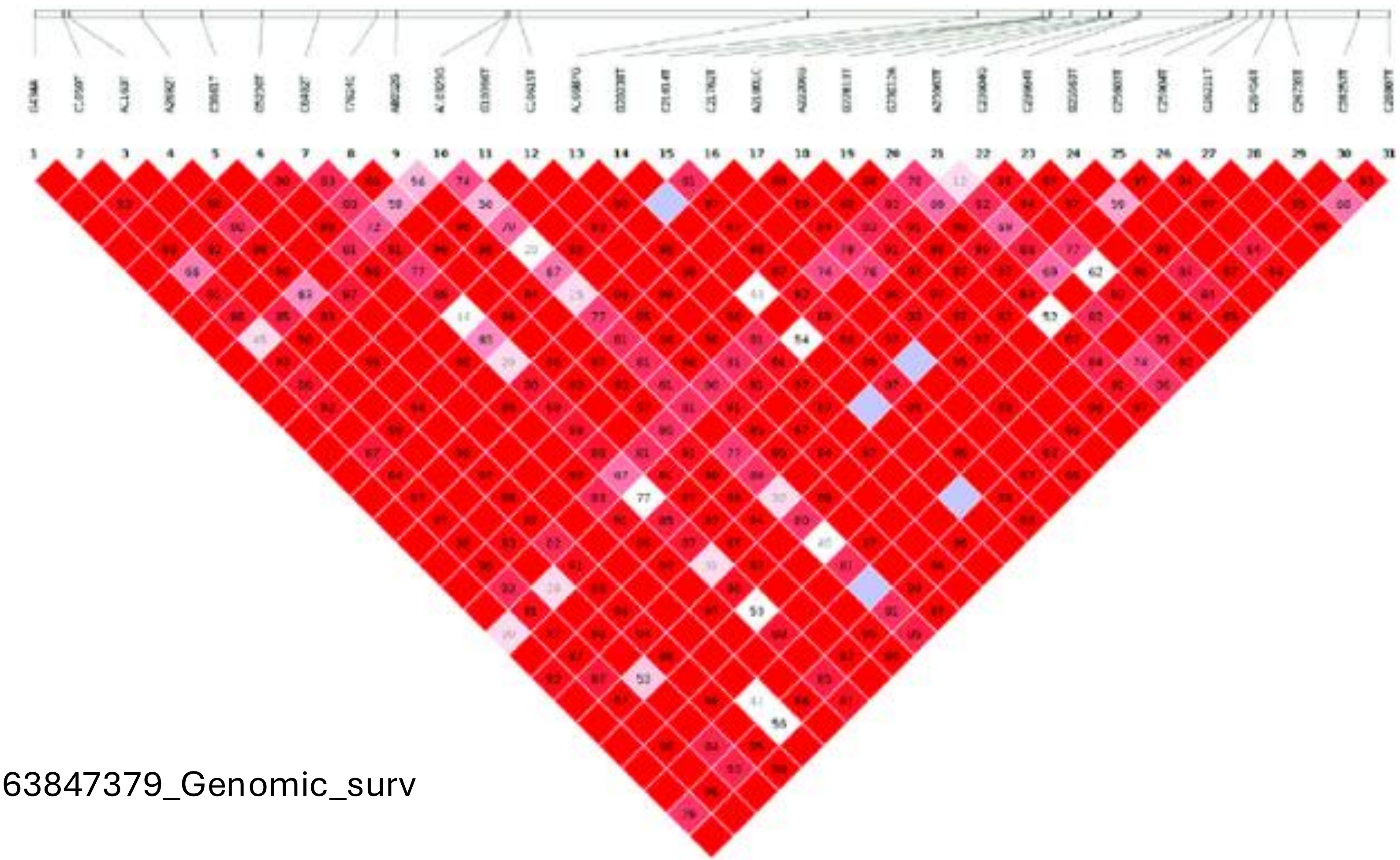
https://software.broadinstitute.org/mpg/sweep/

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

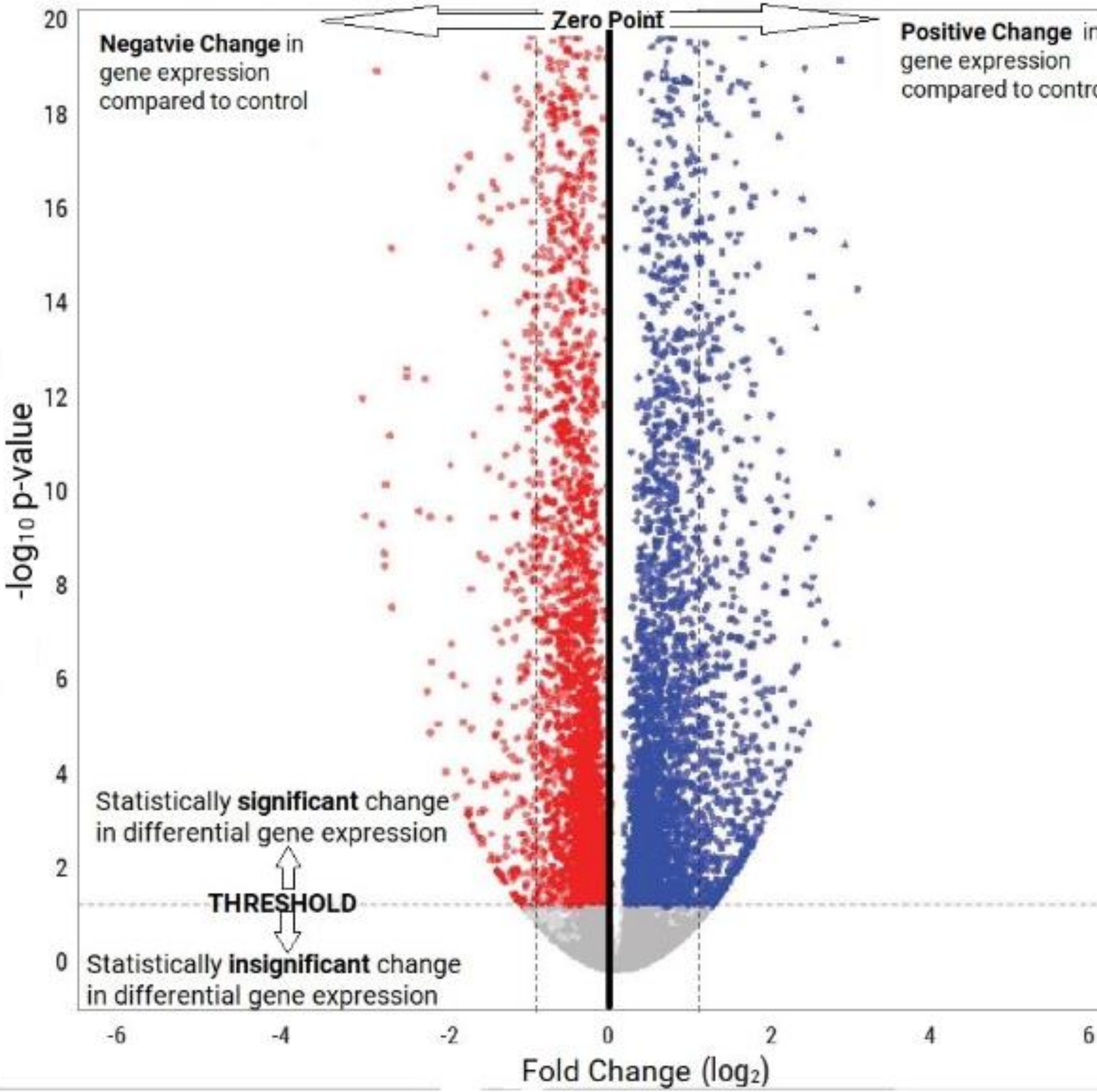


https://www.researchgate.net/publication/363847379_Genomic_surveillance_unfolds_the_SARS-CoV-2_transmission_and_divergence_dynamics_in_Bangladesh

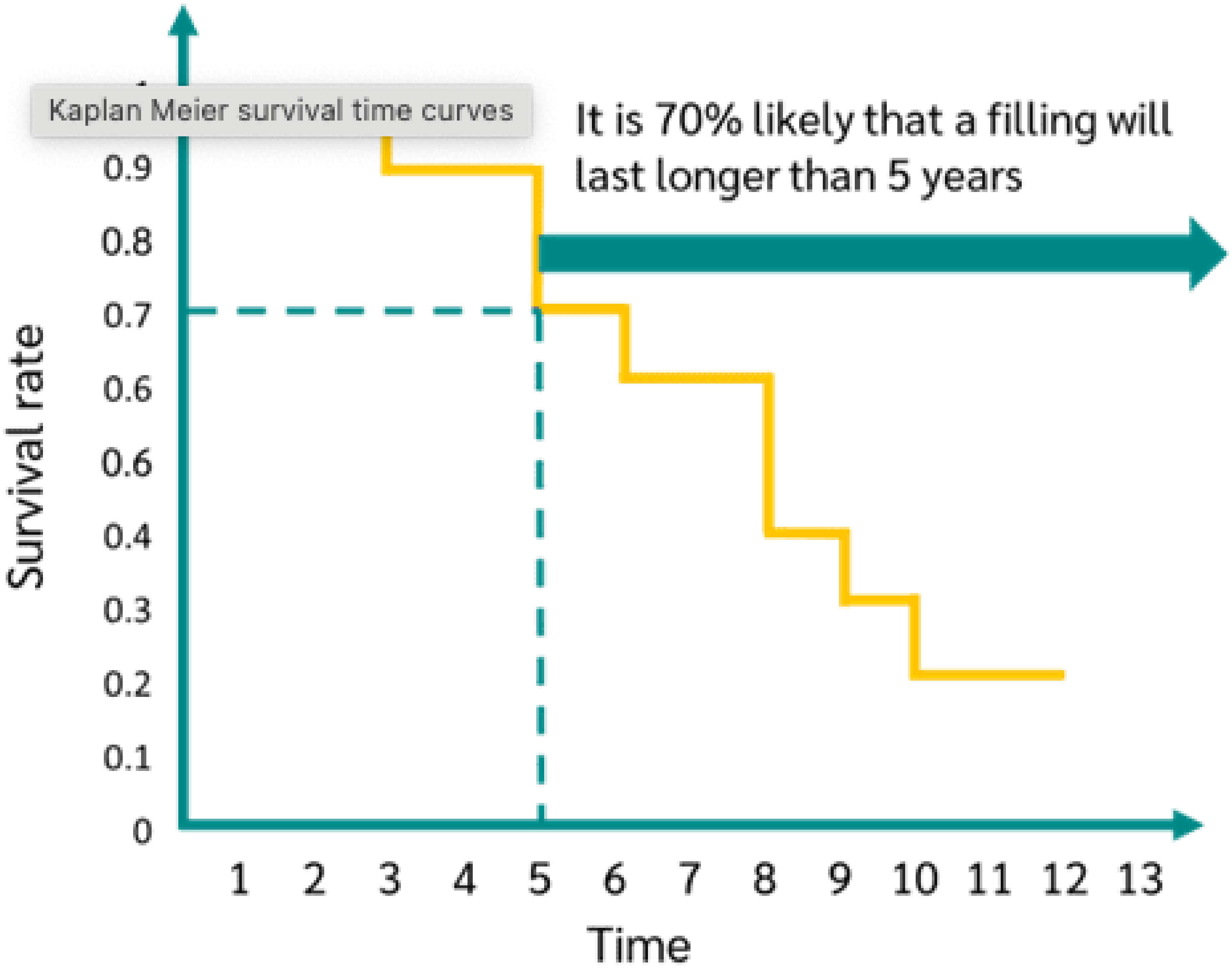
Linkage disequilibrium plot. The LD plot is generated considering the most prevalent SNPs. The number at the top denotes the SNP position, and squares are colored by standard (D'/LOD). The brighter red color indicates a higher D' value and vice versa. The number in square is r 2 value.

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



It is 70% likely that a filling will last longer than 5 years

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”)