# Introduction to Scientific Computing for Biologists

## ISCB20.09 - Introduction to Data Visualization with R
## An Introduction to ggplot2

Md. Jubayer Hossain

https://jhossain.me/

jubayer@hdrobd.org

Founder
Health Data Research Organization
Lead Organizer
Scientific Computing for Biologists

08 February 2021

## What is ggplot2?

- ► ggplot2 is a data visualization package for the statistical programming language R.
- ► Created by Hadley Wickham in 2005, ggplot2 is an implementation of Leland Wilkinson's Grammar of Graphics—a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers.

## Types of Visualization

In statistics, we generally have two kinds of visualization:

▶ Exploratory data visualization: Exploring the data visually to find patterns among the data entities.

▶ Explanatory data visualization: Showcasing the identified patterns using simple graphs.

# Why Visualization?

"A picture is worth a thousand words"

- ▶ Data visualizations make big and small data easier for the human brain to understand, and visualization also makes it easier to detect patterns, trends, and outliers in groups of data.
- ▶ Good data visualizations should place meaning into complicated datasets so that their message is clear and concise.

# Grammar of Graphics

- ▶ Data - The dataset being plotted
- ▶ Aesthetics- The scales onto which we plot our data
- ▶ Geometry - The visual elements used for our data
- ▶ Facet -Groups by which we divide the data

## Data Import

- readr - https://readr.tidyverse.org/
- readxl- https://readxl.tidyverse.org/
- haven - https://haven.tidyverse.org/

# Data Manipulation

- ▶ tidyr- https://tidyr.tidyverse.org/
- ▶ dplyr- https://dplyr.tidyverse.org/

## Data Sources

- https://genome.ucsc.edu/
- https://www.ensembl.org/index.html
- https://www.encodeproject.org/