**SIV Viral Integration Sites and ATAC-seq analysis using R – Getting Started**

Aisha Olufayo (aolufayo@bidmc.harvard.edu) – *Research Intern*, Boudries Lab, CVVR

**Download the workshop package from:** https://github.com/aolufayo-bidmc/IS-ATAC-Workshop, click the Green **CODE** button, and then click **Download ZIP**

Prerequisites:

- R (version 4.4 or later recommended)
- Rstudio (Desktop)

List of R packages Needed: tidyverse, readxl, writexl, openxlsx, readr, BiocManager

You can paste this code in your R console to check you have the required packages and install if necessary:

```
packages <- c("tidyverse", "readxl", "writexl", "openxlsx", "readr", "BiocManager")

for (pkg in packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
}
```

0. ***Must Do:*** open the R project file named "IS_ATAC_Workshop_Files.Rproj"
    - This will put you in the correct working directory for RStudio: access all the R files necessary from the files window of this Rproject
    - Not doing so will result in not being able to access file names, which are written directly in the code for ease
1. **(Cleaning up Integration Site File):** No action, follow along the demonstration
    - Demonstration of the data cleaning process to get integration sites in desired form
    - No need to open any files, you can **skip** this part of running this code by yourself
2. **Annotating SIV Integration Sites:** open and run "2. Annotate Insertion Sites.R"
    - Uses the cleaned data from the previous part to annotate the insertion sites
    - Identifies the corresponding gene of the insertion site, location in gene, sequence, etc.
    - **Output:** "Simulated_IS_annotated.xlsx" in outputs folder
    - If this code does not run, double check "Simulated_IS_clean.xlsx" is in your inputs folder

3. **Overlapping SIV Integration Sites with Open Chromatin:** open and run "3. Alignment of ATAC-seq data with IS data.R"
    - Converts both the IS data and the ATAC-seq peaks to GRanges objects, uses the GRanges function subsetByOverlaps to find integration sites which share at least 1 base pair with open chromatin
    - **Output:** "Simulated_IS_openchr.xlsx" in inputs folder
4. **Annotating IS in Open Chromatin:** open and run "4. Annotating Overlap Sites.R"
    - Takes in the overlaps found in step 3, annotates like in step 2
    - **Output:** "Simulated_IS_openchr_annotated.xlsx" in outputs folder
    - If this code does not run, double check "Simulated_IS_openchr.xlsx" is in your inputs folder