# Introduction to Biocomputing - R Project

**Complete the tasks below and submit your results via a pull request on GitHub by 5 pm Friday, December 3rd.**

You have been contracted by the U.S. Center for Disease Control (CDC) to evaluate the dynamics of an emerging disease outbreak and potential response strategy to the outbreak. Because this outbreak is in the early stages, all details about locations and type of disease will be limited. The scenario is that we have detected, using molecular biology screening, a novel disease-causing agent in two countries. We have extensive screening data made available by both countries and we are interested in answers to two questions:

1. In which country (X or Y) did the disease outbreak likely begin?

2. If Country Y develops a vaccine for the disease, is it likely to work for citizens of Country X?

The molecular screening used is something called "microsatellite analysis". In this specific screen, the presence or absence of ten markers specific to an immunologically active protein from the disease-causing agent give us information about the presense/absence of the disease in a patient and similarity or difference amongst strains of the disease in a group of patients.

- If one or more of the markers are present in a patient's sample, the patient was infected.
- The disease is caused by a bacteria that is transmitted through the air. Given the short generation time of the bacteria causing the disease and rapid sperad of the disease, we suspect the disease-causing bacteria is evolving along its transmission path.
- The gene targeted by the microsatellite screen encodes the main protein shown to form an immunological response by the patient and so differences in which markers are present would indicate differences in the protein in the disease and potentially different responses by a patient's immune system.

*Details about data provided*

Each country is screening a large number of patients with symptoms daily. Data from each country are provided in text files with names of the format "screening_NNN.txt" where NNN indicates the day of year that the screens in that file were conducted on. Each file contains twelve columns - gender, age, and ten columns for the ten microsatellite markers (1 means the marker was present for that patient, 0 means the marker was absent).

Although your primary goal is to answer the two questions above, we expect additional data of the form we currently have access to from Country X and Y in subsequent months. It is also possible that other countries will see the disease spread and therefore we want you to **provide answers and supporting information for the two questions above, but also provide code that could be used for future analyses by the CDC**. Specific requirements of your code are listed below.

*Code requirements*

The CDC requests two scripts written in the R programming language. The first script (*supportingFunctions.R*) will contain a number of custom functions created to accomplish various data handling or summary tasks. The second script (*analysis.R*) will use the `source()` function to load the functions defined in *supportingFunctions.R*, compile all data into a single comma-separated value (`.csv`) file, process the data included in the entire data set in order to answer the two questions above and provide graphical evidence for your answers. Use comments in *analysis.R* to explain the rational and how the graphical evidence supports your answer to the two questions.

To facilitate analysis of the provided data, as well as future data, the CDC requests the following functions (and any others you feel are necessary) be provided in *supportingFunctions.R*:

- Country X and Y have different traditions for the delimiter in their data files. Write a function that converts all files in a directory with space- or tab-delimited data (`.txt`) into comma-separated value files.
- Write a function to compile data from all `.csv` files in a directory into a single `.csv` file. The compiled data should have the original twelve columns from daily data sheets, but also *country* and *dayofYear* columns. The user should be able to choose whether they want to remove rows with NA's in any columns, include NAs in the compiled data but be warned of their presence, or include NAs in the compiled data without a warning
- Write a function to summarize the compiled data set in terms of number of screens run, percent of patients screened that were infected, male vs. female patients, and the age distribution of patients.

You can work in groups of up to 3 students. You will only need to turn in one set of scripts for the whole group via pull request on Github, but all group members must contribute to the final product.

To begin your work, fork the R Project repo from Stuart's Github. Clone the forked repo so that you have the required files. Be sure to commit regularly to show how you and your group members contibuted to your solutions.

## Turning in your assignment via GitHub

Once you have committed all changes to your local Git repo and pushed all of those commits to the forked repo on GitHub, you can "turn in" your assignment using a `pull request`. This can be done from the GitHub repo website. When viewing the forked repo, select "Pull requests" in the upper middle of the screen, then click the green "New pull request" button in the upper right. You'll then see a screen with a history of commits for you and your collaborators, select the green "Create pull request button". In the text box next to your user icon near the top of the page, remove whatever text is there and add "last name submission", but obviously substitute your last names. Then click the green "Create pull request" button.