Rishee Venkatesh

Part 1: Review Questions

General Concepts

1. **What is TCGA and why is it important?**

   The TCGA – The Cancer Genome Atlas – is a catalog of genomic data from over 20,000 samples of 33 cancer types, curated by the National Cancer Institute and the National Human Genome Research Institute. It is a publicly available multi-omic dataset, so it enables researchers to explore a wide range of genes across a large patient sample, which would otherwise be difficult to procure independently. TCGA Data consists of clinical data, mutation data (in MAF format), RNA count data, and protein data through the Clinical Proteomic Tumor Analysis Consortium.

2. **What are some strengths and weaknesses of TCGA?**

   Strengths of the TCGA include the ability to explore a variety of genes across a large patient sample, and the ability to access various different types of data in the form of clinical, mutation, and RNA count data.
   Weaknesses include the amount of NA (unavailable) and missing data for patients, making it incomprehensive and producing limitations on the ability to perform research.

Coding Skills

1. **What commands are used to save a file to your GitHub repository?**

   (upload) git add, git commit -m “message”, git push
   (download) git pull, scp (-r if folder) into personal directory

2. **What command(s) must be run in order to use a package in R?**

   library(), install.packages()

3. **What command(s) must be run in order to use a *Bioconductor* package in R?**

   library(BiocManager)

4. **What is boolean indexing? What are some applications of it?**

   Boolean indexing is the use of Booleans to identify the specific row/column number by setting the value to true when the value matches that of the data frame. Boolean masking is the application of a Boolean vector to a column or row in a data frame. One of the

primary applications of it is to remove NA values from a data frame, or to select data that applies to a certain condition.
ex. na_mask <- !is.na(clinic_rad_merge$person_neoplasm_cancer_status)

5. **Draw a mock up (just a few rows and columns) of a sample dataframe. Show an example of the following and explain what each line of code does.**

   a. **an `ifelse()` statement**

      ifelse(condition, action if true, action if false)

      ifelse (df[2,2] == "G", print("true"), print("false"))

F	**F**	**F**
F	**G**	**F**
F	**F**	**F**

(would print true)

   b. **boolean indexing**

      ex. Looking for 2, 2
      df[col[F, T, F], row[F, T, F]]

F	**F**	**F**
F	**T**	**F**
F	**F**	**F**