**R for bioinformatics**

Instructor: Dmitri Kazmin, Ph.D.

Introduction and class syllabus

**Pre-requisites**. There are no prerequisites as far as coding proficiency, familiarity with R, or knowledge of immunology are concerned. However, there may be a rather steep learning curve, depending on the baseline proficiency level.

**Required software and online access**. Each participant is required to install the latest stable version of R (<https://www.r-project.org>) and R Studio (free desktop version) (<https://www.rstudio.com/products/rstudio/>). I strongly encourage everyone to install the required software ahead of time and familiarize yourself with it prior to the first class.

Also, each participant is required to create an account on Github (github.com), and send me ([dmitri.kazmin@gmail.com](mailto:dmitri.kazmin@gmail.com)) your Github user name so I could invite you to the class repository (<https://github.com/DmitriKazmin/R.class>). This also needs to be done prior to the first class. All codes, source data and other relevant files will be posted on Github. This is a great platform for code exchange and collaborative coding projects.

Also prior to the first class, please create a work directory on your computer to store the codes, source data and output files. Avoid spaces in the pathname! (e.g. use /Python.club instead of /Python club). Download all files from the “Intro” branch in the Github repository to this directory and unzip the archive.

**General class structure**. Science is a practical skill, so is coding. To learn it, you need to practice, and the best way to practice is to do it yourself. No PowerPoint presentations will teach you how to code and how to apply your coding skills to answering biological questions. That being said, we will minimize the PowerPoint presentation part to a bare minimum and will spend most of our time in the R Studio environment, studying the codes that I will provide, figuring out how they work, and creating your own codes based on the provided examples. Generally, each 60-minute class will be split into two 15-minute walk-through/demo/instruction periods alternating with two 15-minute hands-on coding exercises. There will be homework given at the end of each class. If you run into any issues with classwork or homework, please either share your thoughts and ideas via Github wiki, or open an issue on Github under the “issues” tab. If you’re stuck figuring out a coding problem, please share your code when opening an issue. Everyone is encouraged to participate in writing the wiki and in answering open issues. Active participation in the Github activities gets you good karma and everyone will love you for that.

The class will involve the analysis of a real-world data obtained from a real clinical study of the novel influenza vaccine safety and efficacy in young children and adults.

**Syllabus**

**Week 1. Introduction to bioinformatics.** In this class we will cover the following topics: what is bioinformatics and why do we need it? What is the clinical study that we will explore? How do vaccines and adjuvants work? Why do we need to study vaccines and what important information do we hope to obtain from such clinical studies? Finally, we will cover some introductory concepts of programming in R, and, for that matter, in any programming language: what is a variable? What is a function and what are the arguments to a function? We will learn to make our first variables and look at the examples of some basic functions that operate on these variables.

**Week 2. R basics-1.** In this class we will cover different classes of data that R operates with, and different types of variables that can store these data. We will learn to make vectors, matrices, data frames and lists. We will practice manipulating these types of variables, filtering, subsetting, ordering them, and extracting data from complex structures.

**Week 3. R basics-2.** In this class we will focus on more complex examples of R syntax. We will learn to read the data from disk and write our results; we will focus on writing FOR loops and incorporating IF / ELSE statements. By the end of the first three classes, you should be able to write your first functional code to manipulate the input data, create output data structures and save them to disk.

**Week 4. Exploring built-in datasets**. Base R comes with several practice datasets suitable for honing your skills in basic statistical analysis and data visualization. In this class, we will learn how to perform paired and unpaired Student t-test and run ANOVA. We will touch on the subject of parametric vs. non-parametric tests. We will talk about the measures of estimating the significance of the observed effects and correcting the significance of our findings for multiple testing.

**Week 5. Plotting in R.** Base R comes with formidable plotting capabilities. However, to fully take advantage of all the plotting options in R, we typically use dedicated plotting libraries, such as ggplot2. In this class we will learn to make basic types of plots, such as boxplots, scatter plots, histograms and bar plots using ggplot2. We will also learn to make and annotate heatmaps using the pheatmap library.

**Week 6. Gene expression dataset**. In this class we will work with a large dataset encompassing the measurements of changes in gene expression for thousands of genes in the peripheral blood of the study participants. Vaccination causes massive transcriptional responses (that is, changes in gene expression) in circulating blood cells. Which genes are the most affected? By how much? What is the temporal kinetics of transcriptional response? Is the transcriptional response different between children and adults? We will learn to apply basic statistical analyses of gene expression and visualize the results using volcano plots and heatmaps.

**Week 7. Gene set enrichment analysis**. Now that we have measured the transcriptional responses, we need to make the biological sense of the results. Among hundreds or thousands of affected genes, which ones are the most interesting from the biological perspective? Which biological pathways are most affected by vaccination? Which cell types are likely to be the most impacted by the vaccine? To this end, we will apply the Gene Set Enrichment Analysis (GSEA) to the lists of affected genes, and will represent the results as bubble plots and heatmaps.

**Week 8. Antibody responses**. Most vaccines work by stimulating the body to produce large quantities of antibodies against the target pathogen. The influenza vaccine is no exception. In this study we have measured antibody titers against several strains of flu virus prior to the vaccination and a month after. Did all participants respond to vaccination? Were there some that did not respond or responded poorly? Did all participants develop protective immunity against all three vaccine virus strains? Are the observed strain-specific responses to vaccine affected by the pre-existing baseline immunity? Which strain, if any, is best suited to represent the overall vaccine efficacy in each participant? We will spend this class period analyzing the antibody data, and graphically representing our findings to answer the above questions.

**Week 9. Cellular data**. There are many immune cell types that participate in the successful immune response to vaccine or infection: T cells, B cells, macrophages, monocytes, dendritic cells, granulocytes (eosinophils, basophils, neutrophils). Each cell type comes in a wide variety of flavors, that can be distinguished by modern analysis methods. In this class we will study changes in frequency (abundance) of various cellular subsets in response to vaccination, and will graphically represent our findings. We will also investigate whether changes in frequency in any cell population tracks with the accumulation of antibodies, and, perhaps, will come up with a viable biological hypothesis linking the cellular and serological (antibody) responses.

**Week 10. Ouchie time! Vaccine reactogenicity and adverse effects**. As any pharmacological intervention, vaccination carries a risk of adverse effects (AEs). Many of you, who received a COVID-19 vaccine booster have experienced such adverse effects too. Generally, these AEs are mild, such as redness and swelling at the site of the injection, low-grade fever, pain etc. What are the biological mechanisms that determine which subject will develop an adverse effect to the vaccine and which will not? In this class we will combine the gene expression analysis with the AE data to define the biological pathways that track with the development of mild AEs, such as swelling and redness. We will learn to use the overrepresentation analysis (ORA) to define such pathways.

**Week 11. Transcriptional correlates of vaccine efficacy**. In this class, we will utilize the correlation analysis to define the functionally linked groups of genes, whose expression tracks with the accumulation of antibodies against the flu virus, and hence, protective immunity against flu. We will come up with a viable biological hypothesis about the genes that control vaccine efficacy, and will (as always) graphically represent our findings. At the end of the class, we will summarize our findings, and each class participant will be encouraged to write a mini-paper on the transcriptional and cellular correlates of vaccine efficacy and safety.