Bioinformatics Analysis and Visualisation of Medical Genomics Data

KI course number H7F5633, October 2025
Document version date: 2025-08-10

Introduction

This course is based on student feedback such as: *We would like to learn how to make publication-quality plots as they are shown in high-profile genomics journals*. In this course, we invite the lead bioinformatics authors of high-quality genomics articles, ask them to bring the data from their published articles, and guide course participants in redoing some of the published figures. In the mornings we have lectures from the teachers. The afternoons are hands-on workshops where the course participants work in small groups of around five students. The students stay in the same group for the whole week and work on one paper. The bioinformatics lead author of that paper is the mentor for that group.

Learning Outcomes

After the completed course, the participants can *understand* the principles and *perform* basic bioinformatics analysis of genomics sequencing data. The participants can *plan* experimental designs and *critically evaluate* the appropriateness of the different sequencing based omics methods and technologies for genome-wide gene regulation studies.

Practical aspects

This course has two parts:

The first course Week 1 consists of assignments conducted by the course participants from their own work place during Week 1 (September 25 to October 1, 2025) of the course.

The second course Week 2 (October 2 to 8, 2025) will be conducted at RIKEN Yokohama Campus, Japan, and will consist of lectures, discussion and hands-on practical work analyzing datasets using the software package R or alternatively Python.

All course participants are required to address all the below tasks (some tasks are marked as *optional*) and to share the results of assignments with peer-review partners using Gitlab, Github, Fishare or a similar publication platform. Please see submission details and deadlines below.

Course Resources

- The official Karolinska Institutet course announcement is available here:
    - https://doctoralcourses.application.ki.se/fubasextern/info?kurs=H7F5633
- Course material including paper assignments for Week 1, course information, presentations (coming during the course) and data are available in this shared Google Drive folder.

Course part 1 in Week 1

- **September 25 to October 1, 2025**, Online work from home
- Please see instructions below

The 11th RIKEN-KI-SciLifeLab Symposium

We strongly encourage you to attend the The 11th RIKEN-KI-SciLifeLab Symposium Multimodal Fusion and Data Integration in Multi-Omics, Imaging, and Medical Data on **October 1, 2025**, at the RIKEN Yokohama Campus. Please register at the link: https://forms.office.com/r/9u5U1MLKip

Course part 2 in Week 2

- First course day: 2025-10-02, 9:00 Tokyo time
    - Meeting in front of the campus gate of the RIKEN Yokohama Campus, 1-chōme-7-22 Suehirochō, Tsurumi Ward, Yokohama, Kanagawa 230-0045, Japan
- Last course day: 2025-10-08, 18:00 Tokyo time
- Earliest day departure from Japan for participants from Sweden: 2025-10-09

Course location in Week 2

The course is held at the RIKEN Yokohama Campus.
RIKEN access Instructions:
Take the #08 bus from Platform 8 at the East Exit of Tsurumi Station (also accessible from the West Exit of Keikyu Tsurumi Station) and get off at the RIKEN Shidai Daigakuin Mae bus stop. The institute is across the street. All buses from this platform are bound for Fureyu. Buses depart Tsurumi every 5–15 minutes. It takes about 15 minutes to arrive at RIKEN Yokohama. The fare is 240 yen in cash.

Schedule Week 2

The most recent schedule is available in the shared Google Drive folder.

Please note that there are no course events on October 4-5, 2025, (weekend). These days can be used for home studies.

Instructions for course assignments

- Please cover the following aspects in your assignments
    - Show the commands you typed.
    - Include comments in the code explaining what the commands are doing. Usually there is more descriptive text than there is code.
    - Show the responses/ outputs you received from the software after typing the commands; in case of large outputs you can shorten the output.
    - Comment and explain what you learn from the output and what conclusion you can draw.
- Please deposit your assignments in Gitlab, Github, Figshare or similar publication platforms.

Uploading and deadlines for Assignments

Assignment 1

Please, provide your assignment for peer assessment directly to the student you are evaluating (see file “Student Groups.pdf” in folder "Week 1" in this shared Google Drive folder) no later than September 29, 2025, by giving them access to your Gitlab or Github or FigShare (or similar) repository.

Assignment 2

Please, provide your review / feedback directly to the student you are evaluating no later than October 1, 2025.

Course Assignment 1

Task 1 - Literature

[**NOTE:** This task is also the basis of the journal club task on day two of Week 2 (see schedule)]

1. Read the research article of the hands-on working group you are assigned to (see file “Student Groups.pdf” in shared folder *General course material*).
2. Answer the following questions
    a. What is the medically relevant insight from the article?
    b. Which genomics technology/ technologies were used?
3. Further related research questions
    a. List and explain at least three questions/ hypotheses you can think of that extend the analysis presented in the paper.
    b. [Optional] Devise a computational analysis strategy for (some of) the listed questions under 3a.

Task 2 - Git repositories and R Markdown

- Start a new project in a Gitlab, Github or Figshare repository. Check with your doctoral supervisor if you can start a project in the repository of your lab or if you have to start your own repository.
- All documentation of the Assignment 1 has to be provided in your Git/ Figshare repository as a (formatted) text document.
- All documentation of your hands-on work during Week 2 also has to be provided in your repository as R Markdown document(s).

Task 3 - Introduction to R and online R course

- Install the most recent version of the R software on your computer by following the instructions provided at the R software website.
- Install the most recent version of RStudio Desktop (Open Source version) on your computer.
- Bioconductor is an add-on package for R providing tools for the analysis and comprehension of high-throughput genomic data.
  Install the most recent version of the Bioconductor package on your computer.
- Tidyverse is a toolbox for streamlining data import, modeling, transformation, curation, and visualization. Tidyverse tools enable you to make your scripts more reader friendly and overall more neat and efficient.
  https://www.tidyverse.org/packages/
    - Tidyverse has a powerful visualization package called **ggplot2** that we recommend using. With ggplot2, you have more control over your plot's parameters than with the basic R plotting functions.
      https://ggplot2.tidyverse.org/#learning-ggplot2
- [Optional] Conduct the R online course available here:
    - Introduction to R (hands-on exercises, strongly encouraged)
      https://campus.datacamp.com/courses/free-introduction-to-r
      (the course is completely free, but asks after every completed level to sign up for a license, which can be circumvented by going back to the course list)
    - Alternative resources for R
      https://www.datacamp.com/community/open-courses/r-for-the-intimidated
    - R also provides an internal learning package named Swirl. Swirl is interactive and gives feedback on your progress. https://swirlstats.com/

  Depending on your previous experiences with R, taking this R online course might take more or less time. Please make sure to understand the concepts behind the exercises.
- [Optional] You can consider using R-notebooks for Assignment 1. It is up to your preferences.
- Reference for R language
  https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- [Optional] More information on R Markdown and the Markdown cheatsheet
  https://rmarkdown.rstudio.com
  https://raw.githubusercontent.com/rstudio/cheatsheets/master/rmarkdown-2.0.pdf
  (opening the link will download the cheat sheet on your computer)

Task 4 - R basic operations

1. What is the square root of 10?
2. What is the logarithm of 32 to the base 2?
3. What is the sum of the numbers from 1 to 1000?
4. What is the sum of all even numbers from 2 to 1000?
5. How many pairwise comparisons are there for 100 genes?
6. And how many ways to arrange 100 genes in triples?

Task 5 - Using R example datasets

1. Use the R internal CO2 dataset ("data(CO2)").
2. Describe briefly the content of the CO2 dataset using the help function.
3. What is the average and median CO2 uptake of the plants from Quebec and Mississippi?
4. [Optional] In the "airway" example data from Bioconductor, how many genes are expressed in each sample? How many genes are not expressed in any sample?

Task 6 - R Functions

1. Write a function that calculates the ratio of the mean and the median of a given vector.
   This is a helpful measure to detect data with outlying values.
   Note: See *Reference for R language*
2. Write a function that ignores the lowest and the highest value from a given vector and calculate the mean.
3. Read about piping from here:https://r4ds.had.co.nz/pipes.html#pipes (you don't have to learn everything, a basic understanding of the usage is enough). Write a short (max. 300 characters, no spaces) explanation of why, how, and when not to use pipes.
4. Familiarize yourself with the apply-family of functions (apply, lapply, sapply etc.)
   http://uc-r.github.io/apply_family
   Write a short explanation (max. 300 characters, no spaces) of why they could be useful in your work.

Task 7 - Basic visualization with R

Comment: Files are provided in the shared Google Drive folder under *Course week 1*. Examples of how to use visualization functions in R and other highly useful information that may aid you during the course can be found here https://biotechnoalchemist.github.io/Teaching2025/ The nummenmaa-package mentioned within the examples can be installed with the code below if you wish to try some of the examples yourself.

```
install.packages("remotes")
library(remotes)
install_url("http://emotion.utu.fi/wp-content/uploads/2019/11/nummenmaa_1.0.tar.gz",dependencies=TRUE)
```

1. Compare the distributions of the body heights of the two species from the 'magic_guys.csv' dataset graphically
    a. using the basic 'hist' function as well as ‘ggplot’ and ‘geom_histogram’ functions from the ggplot2 package. Optimize the plots for example by trying several different 'breaks'. Note that ggplot2-based functions give you many more options for changing the visualization parameters, try some of them.
    b. Do the same comparison as in a. but with boxplots. If you want to use the ggplot2-package, use the functions ‘ggplot’ and ‘geom_boxplot’.
    c. Save the plots with the ‘png’, ‘pdf’, and ‘svg’ formats. In which situation would you use which file format?
2. Load the gene expression data matrix from the ‘microarray_data.tab’ dataset provided in the shared folder, it is a big tabular separated matrix.
    a. How big is the matrix in terms of rows and columns?
    b. Count the missing values per gene and visualize this result.
    c. Find the genes for which there are more than X% (X=10%, 20%, 50%) missing values.
    d. Replace the missing values by the average expression value for the particular gene. (Note: Imputing data has to be used with caution!)
3. Visualize the data in the CO2 dataset in a way that gives you a *deeper understanding* of the data. What do you see?

Task 8

1. Install the Tidybiology package, which includes the data ‘chromosome’ and ‘proteins’
   `devtools::install_github("hirscheylab/tidybiology")`
    a. Extract summary statistics (mean, median and maximum) for the following variables from the ‘chromosome’ data: variations, protein coding genes, and miRNAs. Utilize the tidyverse functions to make this as simply as possible.
    b. How does the chromosome size distribute? Plot a graph that helps to visualize this by using ggplot2 package functions.
    c. Does the number of protein coding genes or miRNAs correlate with the length of the chromosome? Make two separate plots to visualize these relationships.
    d. Calculate the same summary statistics for the ‘proteins’ data variables length and mass. Create a meaningful visualization of the relationship between these two variables by utilizing the ggplot2 package functions. Play with the colors, theme- and other visualization parameters to create a plot that pleases you.

Course Assignment 2

You are commenting on the *Assignment 1* of a fellow course participant considering the criteria below and send your comments directly to the fellow student.

1. Format of the assignment
    a. Is the assignment well structured?
    b. Is the formatting appropriate?
2. Completeness of the assignment
    a. Are all the tasks addressed?
3. Correctness of the answers
    a. Are the given answers correct to the best of your understanding?
4. Validity of the drawn conclusions
    a. Do the provided analyses address the questions asked?
    b. Are the provided insights conclusive given the conducted analyses?
5. Style of the written R code
    a. Is the provided R code well documented?

Appendix

RIKEN Course Announcement

https://www.ims.riken.jp/english/2025/04/004865.php

Course syllabus KI

Study period: 2025-09-25 - 2025-10-08

Application period: 2025-04-15 - 2025-05-06

Course layout

Week 1: 2025-09-25 to 2025-10-01 Work on task assignments from home.

Week 2: 2025-10-02 to 2025-10-08 on-site with mandatory attendance at RIKEN Yokohama in Japan.

Week 1 of the course is in the form of individual homework assignments (distance course), as preparation for the second week of the course.

Week 2 of the course consists of tasks, lectures, discussions, and seminars in the mornings. In the afternoons, course participants will conduct data analysis in small groups under guidance of a tutor to redo key figures of selected published papers and based on the corresponding published data. The computer language R will be used for the hands-on practical. It is strongly recommended that participants have previous experiences with R.

The course is given in collaboration with RIKEN Yokohama. The course faculty consists of invited speakers from RIKEN Yokohama and from Karolinska Institutet. The course Week 2 takes places in Yokohama, Japan. **Course participants should travel to Japan and arrive in Japan latest 2025-10-01. Departure from Japan should be earliest 2025-10-09.**

Travel and accommodation costs for Swedish students are generally not covered by the course. However, financial support is available for up to ten (10) participants (KI doctoral students/ KI post docs). **Please state clearly in your course application if you are also applying for travel and accommodation support.** Give detailed motivation why you would need the financial support to be able to attend the course. The relevance of your motivation will be the basis to select the funded course applicants. The financial support will thus not be shared equally over all applicants but an amount of approx. 10,000 SEK (including overhead) will be given to the selected participants.

Link to course evaluation

Evaluation link autumn term 2024:

https://survey.ki.se/Report/6QhOTccIn0J

Course description

The purpose of the course is to increase the understanding of the basic principles of bioinformatics and to gain practical skills in bioinformatics analysis of genomic sequencing data using the language R.

Prerequisites and Selection

Prerequisite courses, or equivalent

Intermediate R skills corresponding to course H7F6003 or equivalent.

Selection

Selection will be based on:

1) the relevance of the course syllabus for the applicant’s individual study plan/research (according to written motivation).

2) start date of doctoral studies (priority given to earlier start date).

Course director

Course director: Carsten Daub

Course co-director: Chung Chau Hon from RIKEN IMS, Japan.