Planning CB2040 and BB2250

Alma Andersson

June 16, 2021

Contents

1 Infrastructure

1.1 github

- The github page will be restructured to contain the following three branches:
  - master : instructions with instructions regarding how to download (clone) the correct repo and run the Docker image we will assemble for them.
  - cb2040 : all the content and instructions for the cb2040 students
  - bb2255 : all the content and instructions for the bb2255 students
- By implementing this structure we will minimize the risk to encounter scenarios where students gets confused about which labs they are supposed to work with (which happened last year, although this was due to them not reading the instructions)

- Create a Docker image with all the necessary packages and backend-libraries (system installations) that might be required to execute the labs.

- Person in charge of github chagnes : Alma

1.2 Docker

- Handling Docker images and containers can be a bit messy, so preferably we only want to produce a single image that we distribute by the start of each course session. Therefore, it's **important** that all the labs are finalized at least **one week** before we start so I can compile the image and make sure the containers run smoothly. I also need to be given a list of all packages that are being used in the lab each person is responsible for, both packages that are imported (`library(package)`) as well as packages from which specific functions are imported by namespace calling (`package::fun`).

- We'll be using the a rocker extension where rstudio is included, see `https://github.com/rocker-org/rocker`. This should give a fairly seamless user-experience for the students, as long as they manage to install Docker on their computers. This allows the students to run rstudio in their web-browser (locally hosted).

- The students will all be working with the same image, but hosting their own containers (as one does with Docker). However, there is one big caveat to using these transient containers, being that if they are removed or damaged all their data/progress is lost - also, it's often hard (for someone inexperienced) to access the files/directories inside a container from the host-system; making it troublesome for them to upload their pdf-files. I've evaluated some different options, and believe the easiest way to circumvent this issue is to use so called bind-mounts (see `https://docs.docker.com/storage/bind-mounts/`). The bind-mounts are easier to handle than volumes and tmpfs mounts are (obviously) only active as long as the container is running.

  The students wouldn't have to do any extra-work with the bind mounts, attaching them to the containers is something that will be included in the - by us - provided in the `docker run...` command used to create containers.

- Person in charge of Docker-related work: Alma

2 Exercises

- In general we will try to harmonize the exercises to create a more continuous narrative, where the students focus on a single tissue type in Lab 2-4 (Lab 1 is the introduction to R and does not include any analysis).

- The tissue type of interest will breast cancer data, which is preferable for several reasons, two of them being:

    1. Data accessibility : there are plenty of public bulk RNA-seq, single cell and Visium data available, which can be easily accessible and used in our labs.
    2. Relevancy : it's easier to engage the students in any form of work if the questions they are working with; casting the exercises as analysis and characterization of a well-known disease will hopefully have a positive effect on interest.

2.1 Lab 1

- Lab 1 will remain more or less unchanged as it focuses on an introduction to R. Only minor changes w.r.t. language and grammar will be adjusted.

- Person in charge of revising Lab 1 is : Alma

- Person in charge of grading Lab 1 : Alma

2.2 Lab 2

- Following the discussion Lab 2 will be revised putting more emphasis on bulk RNA-seq, and less focus on the GWAS part.

- My suggestion is to use one of the breast cancer data sets that can be found in the TCGA (The Cancer Genome Atlas). This data has plenty of meta-data associated with it, allowing us to conduct analyses that produces Kaplan-Meier plots representing the survival curves of different strata etc. This also, to some extent, highlights how it's currently more feasible to collect bulk RNA-seq from a large cohort of patients, while single cell and spatial RNA-seq are often more commonly used for in-depth analysis of a small set of patients. The idea would also

be to highlight some of the weaknesses with bulk RNA-seq, setting the stage for Lab 3 (single cell RNA-seq).

- Person in charge of rewriting Lab 2 : Alma

2.3 Lab 3

- Lab 3 needs some minor updates, to be more consistent with the new narrative (breast cancer focus). This would mainly include:
  - Change of data set to work with, proposed data is Alex Swarbrick's HER2 data.
  - Clearly highlighting the differences between single cell RNA-seq and bulk, i.e., putting emphasis on what sets them apart and what information that can be gained from single cell RNA-seq (hopefully mention intraatient heterogeneity).
  - Also clearly highlighting what information we can't obtain from the single cell RNA-seq study, but which spatial transcriptomics may provide (setting the stage for Lab 2).
- This lab is somewhat tricky in the sense that it should be standalone from Lab 2 (since students of BB2255 aren't doing that lab), but also fit into the story.
- Person in charge of rewriting Lab 3 : Sami

2.4 Lab 4

- Lab 4, just as Lab 3, will not have to go through any major changes, but just update the data set that we will be working with.
- As a suggestions we can work with the 10x publicly available breast cancer data sets.
- Ideally we will also map the single cell data (used in Lab 3) onto the spatial transcriptomics data, which closes the story in a neat way.
- Important is that we highlight how the two different modalities (single cell vs. spatial) complement each other.
- Person in charge of rewriting Lab 3 : Alma