# Episode 3. Tools for Oracles and Overlords

## Jupyter notebooks for reusable data analysis

### Exercise 1: Basics of Jupyter Notebooks

Open this Jupyter Notebook: <http://mango.bio.ed.ac.uk/jupyter>

We will first show you how to duplicate a notebook, save it and run code.

1.     Select the notebook titled **'student\_notebook\_light\_conditions.ipynb**' and click 'Duplicate'.

2.    Confirm with **Duplicate** when asked if you are certain that you want to duplicate the notebook.

3.    A copy of the notebook has appeared with the suffix '-Copy' and a number.

4. Rename the notebook from -copy number to your initials e.g. “student\_notebook\_light\_conditions\_TZ” , open it by clicking

Explore the anatomy of the notebook (<https://github.com/carpentries-incubator/fair-bio-practice/blob/gh-pages/fig/10-02-jupyter_anatomy.png> )

5.    Save the notebook: Click on the disk symbol in the toolbar

6.    Run the notebook: Select the top cell of the notebook with the title (this is likely pre-selected already and will show with a light-blue bar to its left), and click “Run” in the tool bar.

7.    Click two times. What can you see?

-

-

8.    We want to run ALL the code: In the top tool bar click Cell > Run All. What can you see?

-

-

 DONE:

 ------------------------------------------------------------------------------------------------------------------

### Exercise 2: How to add and remove content

You have duplicated a notebook and saved it under your own name. Now we will add text, remove cells and change code.

1.     Change the author name of the document to your name: Double click on the cell containing the author name and change the name.

2.     Press Run again.

3.     You can experiment with formatting, dbl click on the cell,   
add # or ## or ### before text, surround text with \*text\* \*\*text\*\* ; Run the cell again

4.     Change colours of your graph: Where the code of the graph reads the comment “# change colour of groups” you can replace the HEX codes, # followed by 6-symbol code, with names of colours (e.g. blue, green…) or other HEX codes if you are familiar with them.

8.     Save the graph under new name: Add your initials to the file name under which the image is saved. Press Run. Your image should be visible in overall file hierarchy.

DONE:

 ------------------------------------------------------------------------------------------------------------------

### Exercise 3: Sharing of your Jupyter Notebook

You have now generated your own analysis and want to share this with your colleagues.

1. Download your Notebook (ensure all code has been run) as .html

-----------------------------------------------------------------------------------------------------------

### Exercise 4: Accessibility of Jupyter Notebooks

On a scale from -2 to 2, based on these examples how do you feel about the following statements (R is interchangeable with Python),

where -2 (strongly disagree), 0 no opinion to +2 strongly agree:

Type a number between -2 to 2 behind the statement:

• it is easier to generate a series of plots with similar layout in R than Excel:

• using notebooks does not require any programming knowledge:

• notebooks give you a better overview of your data analysis than Excel:

• notebooks replicates laboratory style records for data analysis:

• you need to learn R to do any data processing in notebooks:

• notebooks assures reproducible computing:

DONE:

-----------------------------------------------------------------------------------------------------------

GitHub

<https://github.com/ewallace/pseudonuclease_evolution_2020/commits/master>

-----------------------------------------------------------------------------------------------------------

Benchling

<https://benchling.com/s/etr-8u10BD9KalyiIh9ASHXA?m=slm-IOyleZlGh3xMpJXoWRrO>

https://benchling.com/covid\_ww/f/lib\_tpnkCXlW-archived-experiment-notes/etr\_92p2OLL5-pcr004-test-of-existing-rt-qpcr-probes-and-primers-on-cww-samples/edit

-----------------------------------------------------------------------------------------------------------

## Exercise 5. Tools and the management

Imagine your team has successfully adopted ELN , Jupyter notebooks and version control with GIT for your data. You are facing one of the scenarios bellow. Based on what you learnt from the short demonstrations, as well as your “breakout room” experience, how these tools could help you to act in the assigned situation. What actions would you take?

For example:   
SCENARIO: Your building has been flooded, all electronic equipment has been damaged, you won’t have access to your office nor lab for the next 3 months.

REACTION: You mostly worried about the delay and loss of biological samples. Your data and methods are safe on the servers.

ACTIONS: - you ask group members to list experiments which needs repeat

* You divide group into those who have data to analysed and can work on publications, and those who will spend the next couple of months on training and learning new programming skills

Room 1:

You suddenly lost a postdoc because she/he has won the National Lottery and won’t be coming to work any more. They will stay two more weeks with you, what actions could you take:

-

-

-

-

-

-

Room 2

Your new postdoc has been trying to setup RT-qPCR for strains detection of Staphylococcus aureus for the last 3 months. He says he has no success and looks super stressed and depressed. Your group previously successfully used RT-qPCR for strains detection of E.Colie. What actions could you take:

-

-

-

Room 3

Your new postdoc has been trying to setup RT-qPCR for strains detection of Staphylococcus aureus for the last 3 months. He says he has no success and maybe next month. What worries you is that you never bumped into him in the lab whenever you happen to come by with a visit. Your group previously successfully used RT-qPCR for strains detection of E.Colie. What actions could you take:

-

-

-

Room 4

You got a teaching grant for your microscopy facility. You will have at least 12 visiting ECR a year all learning how to use single-molecule fluorescence microscopy to quantify both transcripts and protein levels. You would like to take advantage of this “free work power” and use them for your drug screening experiments. What actions could you take:

-

-

-

## Exercise 6. Learning curve

What you is the minimum knowledge that you as a PI should know about each of these tools to have them adopted in your team? Or what should be the basic skill set for the PI with those tools.

For example:

ELN – know which one is being used by the group and why this one

ELN

-

-

-

Git/GitHub

-

-

-

Jupyter Notebooks / R-markdown

-

-

-

## Feedback: Tools lesson

On the scale 0 - 5 (zero a terrible lesson, 5 a fantastic lesson)

How good was this lesson:

-

On the scale 0 - 5 (zero not at all, 5 yes it was productive way of spending my time)

Was it worth your time:

-

-----------------------------------------------------------------------------------------------------------

# Episode 4. Public repositories

## Exercise 1: Public general record (demo 4 mins + 4 mins evaluate + 4 mins finding with instructor)

Have a look at the following record for a data set in Zenodo repository:

Boehm et al. (2020). Confocal micrographs and complete dataset of neuromuscular junction morphology of pelvic limb muscles of the pig (Sus scrofa) [Data set]. In Journal of Anatomy (Vol. 237, Number 5, pp. 827–836). Zenodo.

<https://doi.org/10.5281/zenodo.5045374>

Let’s check which elements make it FAIR? (instructor should explain the elements)

FINDABLE (persistent identifiers, easy to find data and metadata):

ACCESSIBLE (The (meta)data retrievable by their identifier using standard web protocols):

INTEROPERABLE (The format of the data should be open and interpretable for various tools):

REUSABLE (data should be well-described so that they can be replicated and/or combined in different settings, and any conditions on reuse stated clearly with a licence):

Exercise 1a (4min)

Now, skim through the data set description (HINT there is also a README), try to judge the following, and indicate your evaluation using marks from 0 to 5 (5 best) as to whether:

• It is clear what the content of the data set is:

• It is clear why the data could be used (ie what for):

• It is well described:

• How confident will you be to work with this data set:

• How easy it is to access the data set content:

• Your team datasets are equally well described (or better):

DONE:

------------------------------------------------------------------------------------------------------------------

## Exercise 1b: Dataset discovery

Try to find:

- data sets related to neuromuscular junction in Zenodo

Judge the following, indicating your assessment using marks from 0 to 5 (5 best)

• how easy it is to find similar or interesting data sets:

• It is clear what the content of the other data sets are:

• It is clear why the data could be used (ie what for):

• They are well described:

DONE:

------------------------------------------------------------------------------------------------------------------

## Exercise 2: Domain specific repositories (5 min)

Select one of the following repositories based on your expertise/interests:

Have a look at mRNAseq accession 'E-MTAB-7933' in ArrayExpress

(https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-7933/)

• What makes it better than Zenodo?:

• What domain specific features can you see?:

• Searching:

Have a look at microscopy 'project-1101' in IDR

(https://idr.openmicroscopy.org/webclient/?show=project-1101)

• What makes it better than Zenodo?:

-

-

-

• What domain specific features can you see?:

-

-

-

• Searching:

-

-

-

Have a look at the synthethic part record 'SubtilinReceiver\_spaRK\_separated' within the 'bsu' collection in SynBioHub (https://synbiohub.org/public/bsu/SubtilinReceiver\_spaRK\_separated/1)

• What makes it better than Zenodo?:

-

-

-

• What domain specific features can you see?:

-

-

-

• Searching:

-

-

-

Have a look at the proteomics record 'PXD013039' in PRIDE

(https://www.ebi.ac.uk/pride/archive/projects/PXD013039)

• What makes it better than Zenodo?:

• What domain specific features can you see?:

• Searching:

Have a look at the metabolomics record 'MTBLS2289' in Metabolights (https://www.ebi.ac.uk/metabolights/MTBLS2289/descriptors)

• What makes it better than Zenodo?:

• What domain specific features can you see?:

• Searching:

DONE:

------------------------------------------------------------------------------------------------------------------

## Exercise 4: Finding a repository (3 min +3)

Our own curated repository list:

<https://www.wiki.ed.ac.uk/display/RDMS/Suggested+data+repositories>

Using Fairsharing (<https://fairsharing.org/>) find a repo for flow cytometry data and type the name below:

-

-

-

once done, search for repository for genomics data

-

-

-

DONE:

-----------------------------------------------------------------------------------------------------------

## Exercise 5: Using repositories (5)

Describe/Type:

• What’s your favourite research data repository? Why?

• How can selecting a repository for your data as soon as an experiment is performed (or even before!) benefits your team research and helps data become FAIR?

• What to do if your publication contains multiple data types?

DONE:

## Feedback: Repositories lesson

On the scale 0 - 5 (zero a terrible lesson, 5 a fantastic lesson)

How good was this lesson:

-

On the scale 0 - 5 (zero not at all, 5 yes it was productive way of spending my time)

Was it worth your time:

-

-----------------------------------------------------------------------------------------------------------

# Episode 5. It is all about planning

## Exercise 1: OS and Data Management practices

Read through the following activities / practices, type next to each

-1 if your team/group do not perform it

0 if you are not completely sure what it stands for

+1 if your team/group adheres / practices it

* include license with datasets:
* include license with code / scripts:
* use git as version control:
* store code in github:
* create DOI for datasets / code:
* add date availability section to a manuscript:
* use minimal information standards:
* use ontology terms:
* use generic data repository:
* use domain specific data repository:
* have description templates for various techniques in the lab:
* store data in a shared, network drive:
* have an automatic backup solution for files:
* follow a file naming conventions:
* create standard project folder structure:
* use Electronic Lab Notebooks:
* create figures and plots in python/R:
* select data repository:
* know non-restrictive licenses:
* create readme for each dataset:
* use institutional repositories:
* use controlled vocabularies:
* have ORCID
* have dedicated folder / database for protocols / SOP
* have a way to reference different versions of a protocol
* convert numerical data to csv:
* follow conventions for tidy data tables:
* use jupyter notebooks or R-markdown:
* use metadata format / standards:
* use PID from repositories (eg UniProt, GenBank) in data description:
* use database for bio-samples / strains etc:
* use pipelines for data analysis:
* can access all group data from your own PC:
* discuss with team how particular data type should be described and stored:
* use tools / resources you organization offers for data management:
* use support you organization offers for data management:

DONE:

-----------------------------------------------------------------------------------------------------------

## DMP example and reusable paragraphs

<https://www.wiki.ed.ac.uk/x/yesNGQ>

-----------------------------------------------------------------------------------------------------------

## Exercise 2: Data Management Plan

In your group evaluate the example data management plan. Mark / Comment any section which should be improved.

As a “grant reviewer” focus on following aspects:

* does it look as if the author(s) thought what they would do with their data or is it some ‘whatever’ copy pasted text
* does it look as if the data will be stored in a secured way
* does it look as if the data will be shared in a FAIR way
* are there any standards that should be followed
* does it look as if the data will be made accessible / findable to others for re-use
* are the usage restrictions justifiable

-----------------------------------------------------------------------------------------------------------

## Feedback: DMP

On the scale 0 - 5 (zero a terrible lesson, 5 a fantastic lesson)

How good was this lesson:

-

On the scale 0 - 5 (zero not at all, 5 yes it was productive way of spending my time)

Was it worth your time:

-

-----------------------------------------------------------------------------------------------------------

# Episode 6. Wrapping it up

## Exercise 1: Adoption of Data Management

Select 3 actions that you believe have the biggest impact in improving productivity and data management of your group, type +1 next to them

Check (meta)data standards for the experimental data in your group

Investigate domain specific repositories for your domain(s)

Mandate (FAIR) data management training for your group

Introduce common folder structures for the projects

Develop templates for experiment types

Encourage data plotting/analysis in R/Python

Create catalogue of group research outputs

Use shared network (or cloud) storage for group data

Introduce ELN

Collect group protocols in shared/online platform

Introduce version control (git) for scripts and analysed data

Adopt Jupyther notebook (R-markdown) for data analysis

Require DMP for each project

Develop strain / samples data base

-----------------------------------------------------------------------------------------------------------

## Exercise 2: Policy enforcement without micromanagement

Propose solution how you can enforce / propagate good data management practices without excessive micromanagement and policing.

-

-

-

-

-

-

-----------------------------------------------------------------------------------------------------------

## Feedback: 2nd Session, tools, repositories and DMP

1. How do you feel about the presented topics after this session (type +1 next to the statement that best describes your feeling):

• I am more confused:

• I have a better understanding of them now:

• My knowledge has not changed much:

2. How was the pace of the lessons:

• Too fast:

• About right:

• Too slow:

## Feedback For **The Whole Course**:

On the scale 0 - 5 (zero a terrible course, 5 a fantastic course)

How good was this course:

-

On the scale 0 - 5 (zero useless, 5 useful)

How useful was this course:

-

On the scale 0 - 5 (zero only for my worst enemy, 5 highly recommended)

How likely are you to recommend this course to other PIs:

-

On the scale 0 - 5 (not chance, definitely will send)

How likely are you to send your minions to the 2-day full course “FAIR in bio practice”, which follow similar format to this one:

-

On the scale 0 - 5 (zero not at all, 5 yes it was productive way of spending my time)

Was it worth your time:

-

-

What other topics would you add:

-

-

What topics could we shorten:

-.

-

Would you like this course to be 1.5 longer but have practical excercises with ELN and Version Control? (T F)

-